Tampere University

SHARATH ADAVANNE

Sound Event Localization, Detection, and Tracking by Deep Neural Networks

Tampere University Dissertations 219

Tampere University Dissertations 219

SHARATH ADAVANNE

Sound Event Localization, Detection, and Tracking by Deep Neural Networks

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for public discussion in the Auditorium TB214
of the Tietotalo, Korkeakoulunkatu 1, Tampere,
on 4th March 2020, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos	Dr. Tuomas Virtanen Tampere University Finland	
Supervisor	Dr. Tuomas Virtanen Tampere University Finland	
Pre-examiners	Dr. ir. Emanuël Habets International Audio Laboratories Erlangen Germany	Dr. Hannes Gamper Microsoft Research USA
Opponents	Dr. ir. Emanuël Habets International Audio Laboratories Erlangen Germany	D.Sc. (Tech) Toni Hirvonen Yousician Finland

Cover design: Roihu Inc.

ISBN 978-952-03-1461-3 (print)
ISBN 978-952-03-1462-0 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1462-0

PunaMusta Oy – Yliopistopaino
Tampere 2020

Abstract

In this thesis, we present novel sound representations and classification methods for the task of sound event localization, detection, and tracking (SELDT). The human auditory system has evolved to localize multiple sound events, recognize and further track their motion individually in an acoustic environment. This ability of humans makes them context-aware and enables them to interact with their surroundings naturally. Developing similar methods for machines will provide an automatic description of social and human activities around them and enable machines to be context-aware similar to humans. Such methods can be employed to assist the hearing impaired to visualize sounds, for robot navigation, and to monitor biodiversity, the home, and cities.

A real-life acoustic scene is complex in nature, with multiple sound events that are temporally and spatially overlapping, including stationary and moving events with varying angular velocities. Additionally, each individual sound event class, for example, a *car horn* can have a lot of variabilities, i.e., different cars have different horns, and within the same model of the car, the duration and the temporal structure of the horn sound is driver dependent. Performing SELDT in such overlapping and dynamic sound scenes while being robust is challenging for machines. Hence we propose to investigate the SELDT task in this thesis and use a data-driven approach using deep neural networks (DNNs).

The sound event detection (SED) task requires the detection of onset and offset time for individual sound events and their corresponding labels. In this regard, we propose to use spatial and perceptual features extracted from multichannel audio for SED using two different DNNs, recurrent neural networks (RNNs) and convolutional recurrent neural networks (CRNNs). We show that using multichannel audio features improves the SED performance for overlapping sound events in comparison to traditional single-channel audio features. The proposed novel features and methods produced state-of-the-art performance for the real-life SED task and won the IEEE AASP DCASE challenge consecutively in 2016 and 2017.

Sound event localization is the task of spatially locating the position of individual sound events. Traditionally, this has been approached using parametric methods. In this thesis, we propose a CRNN for detecting the azimuth and elevation angles of multiple temporally overlapping sound events. This is the first DNN-based method performing localization in complete azimuth and elevation space. In comparison to parametric methods which require the information of the number of active sources, the proposed method learns this information directly from the input data and estimates their respective spatial locations. Further, the proposed CRNN is shown to be more robust than parametric methods in reverberant scenarios.

Finally, the detection and localization tasks are performed jointly using a CRNN. This method additionally tracks the spatial location with time, thus producing the SELDT

results. This is the first DNN-based SELDT method and is shown to perform equally with stand-alone baselines for SED, localization, and tracking. The proposed SELDT method is evaluated on nine datasets that represent anechoic and reverberant sound scenes, stationary and moving sources with varying velocities, a different number of overlapping sound events and different microphone array formats. The results show that the SELDT method can track multiple overlapping sound events that are both spatially stationary and moving.

Preface

This study was carried out at Tampere University (previously known as Tampere University of Technology), Finland between 2016 and 2019.

First and foremost, I would like to sincerely thank my supervisor Tuomas Virtanen for his invaluable advice and guidance in my research. I am also grateful for the time and effort he has spent on discussing ideas and revising articles.

I would like to thank all my colleagues at the Audio Research Group. My research and life have both benefited a lot from the innumerable discussions we have had.

Finally, I would like to thank my father Shivaprakash, mother Girija, and wife Jyotsna for supporting me unconditionally.

Contents

Abstract i

Preface iii

Acronyms vii

Nomenclature ix

List of Publications xi

1 Introduction 1
- 1.1 Sound Event Localization, Detection, and Tracking . . . 1
- 1.2 Objectives of the Thesis . . . 2
- 1.3 Main Results of the Thesis . . . 2
- 1.4 Outline and Structure of the Thesis . . . 4

2 Problem Description 5
- 2.1 Problem Formulation . . . 5
- 2.2 Applications . . . 6
- 2.3 Challenges . . . 7
- 2.4 Evaluation . . . 9

3 Background 17
- 3.1 Sound Representation . . . 17
- 3.2 Machine Learning . . . 18
- 3.3 Deep Neural Networks - Architecture . . . 19
- 3.4 Deep Neural Networks - Learning . . . 21
- 3.5 Recurrent Neural Network . . . 23
- 3.6 Convolutional Neural Network . . . 24
- 3.7 Convolutional Recurrent Neural Network . . . 25

4 Sound Event Detection 27
- 4.1 Related Works . . . 27
- 4.2 Sound Event Detection in Multichannel Audio Using Spatial And Harmonic Features . . . 28
- 4.3 Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network . . . 31
- 4.4 Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features . . . 37

5 Sound Event Localization, Detection, and Tracking **43**
5.1 Related Work 43
5.2 Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network 45
5.3 Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Network 50
5.4 Localization, Detection, and Tracking of Multiple Moving Sources with Convolutional Recurrent Neural Network 58

6 Conclusions and Future Work **63**
6.1 Conclusions 63
6.2 Future Work 64

Bibliography **67**

Publications **77**

Acronyms

ASR	automatic speech recognition
CE	cross entropy
CNN	convolutional neural network
CRNN	convolutional recurrent neural network
DFT	discrete Fourier transform
DOA	direction of arrival
DNN	deep neural network
ER	error rate
FC	fully-connected
FFT	fast Fourier transform
FOA	first-order ambisonics
GMM	Gaussian mixture model
GRU	gated recurrent unit
HMM	hidden Markov model
IR	impulse response
LSTM	long short-term memory
MSE	mean square error
MSET	multiple sound event tracking
MUSIC	multiple signal classification
NMF	non-negative matrix factorization
ReLU	rectified linear unit
RNN	recurrent neural network
SEC	sound event classification
SED	sound event detection
SELD	sound event localization and detection
SELDT	sound event localization, detection, and tracking
SGD	stochastic gradient descent
SNR	signal-to-noise ratio
STFT	short-time Fourier transform
SVM	support vector machines

Nomenclature

Latin alphabet

x	scalar
$\mathbf{x}$	vector with entries x_i
$\mathbf{X}$	matrix with entries $X_{i,j}$
$\mathbf{x}_t$	vector representing the column t of matrix $\mathbf{X}$
$\mathcal{X}$	tensor/array with three or more dimensions $\mathcal{X}_{i,j,k,\ldots}$

List of Publications

This thesis contains the following six publications referred as [I] to [VI]. The original publications are reproduced with permission from the respective copyright holders.

I Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, Tuomas Virtanen, "Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features," *Detection and Classification of Acoustic Scenes and Events (DCASE).* Budapest, Hungary, pp. 6-10, September 2016.

II Sharath Adavanne, Pasi Pertilä, Tuomas Virtanen, "Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* New Orleans, USA, pp. 771-775, March 2017.

III Sharath Adavanne, Archontis Politis, Tuomas Virtanen, "Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features," *International Joint Conference on Neural Networks (IJCNN).* Rio de Janeiro, Brazil, pp. 1-7, July 2018.

IV Sharath Adavanne, Archontis Politis, Tuomas Virtanen, "Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network," *European Signal Processing Conference (EUSIPCO).* Rome, Italy, pp. 1462-1466 , September 2018.

V Sharath Adavanne, Archontis Politis, Joonas Nikunen, Tuomas Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Network,"*IEEE Journal of Selected Topics in Signal Processing (JSTSP).* Volume 13, Issue 1, pp. 34-48, March 2019.

VI Sharath Adavanne, Archontis Politis, Tuomas Virtanen, "Localization, Detection, and Tracking of Multiple Moving Sources with Convolutional Recurrent Neural Network,"*Detection and Classification of Acoustic Scenes and Events (DCASE).* New York, USA, pp. 20-24, October 2019.

Author's contributions

Sharath Adavanne is the main author of all the publications in this Thesis. Tuomas Virtanen supervised the work for all the publications, in terms of discussing ideas, and giving feedback on experiments and documentation.

The initial idea for [I, II] was proposed by Tuomas Virtanen, and for [III-VI] was proposed by Sharath Adavanne. The software implementation and experiments have all been

conducted by Sharath Adavanne. Archontis Politis assisted with dataset creation and baseline implementation in [IV] for which he is an equal main author. Archontis Politis has also helped in creating datasets, advised and given feedback for [III-VI]. A part of datasets used in [V] were contributed by Joonas Nikunen. Pasi Pertilä assisted with the code in [I, II]. Toni Heittola and Giambattista Parascandolo assisted with code in [I].

1 Introduction

Smart electronic devices have become an integral part of human life in recent years. This human-device interaction is only going to become more seamless and deeply interconnected in the future. One of the dimensions for interaction can be enabled through machine audition, which will allow devices to understand and interact naturally with the acoustic scene around it. This interaction can be as simple as a smartphone detecting the ambient noise around it and adjusting the ringtone volume automatically; or as complex as a robot recognizing the voice of a moving talker among other sound events in the acoustic scene, and further interact and navigate with the speaker. This property of a device to automatically recognize and understand the events happening around it without human input is called context awareness [1]. The applications of such machine audition are limitless and will lead to smarter context-aware devices.

Humans have evolved to recognize and localize different sound events around them using just the audio. They can perform this recognition and localization in an acoustic scenario with multiple sound events overlapping both temporally and spatially. But for the current devices, this is a challenging task. Recognition and localization of sounds are what makes humans audio-context-aware, and teaching machines to do this will bring machines one step closer to human audition. Such an audio-context-aware device, for example, can assist hearing-impaired people to visualize sound events. In this regard, the focus of this thesis is to develop methods that are inspired by the human auditory system to track the spatial trajectory and further recognize multiple sound events in a dynamic acoustic scene.

1.1 Sound Event Localization, Detection, and Tracking

A sound event is a segment of audio that can be labeled as a distinctive concept [2]. Some common sound event classes that we come across in our everyday lives are bird calls, dog bark, speech, and music. In a real-life acoustic environment, these sound events do not occur in isolation but often overlap with each other temporally. The task of recognizing such overlapping sound events and detecting their corresponding onset and offset times is often referred to as sound event detection (SED).

Sound event localization can be defined as the task of determining the direction or the position of an acoustic source with respect to the microphone. Estimating only the direction of the source around the microphone is referred to as the direction of arrival (DOA) estimation and is often represented using the horizontal (azimuth) and vertical (elevation) angle of a 2D spherical coordinate system. In a real-life acoustic environment, in addition to temporal overlapping, sound events also overlap spatially with each other. Further, sound events can be both spatially stationary and moving with varying angular velocities. The task of detecting the individual DOAs of such overlapping sound events

and tracking them over time is referred to as multiple sound event tracking (MSET). Finally, the combined task of SED and MSET is referred as SELDT, i.e., temporally detecting the onset and offset times of each sound event, localizing and tracking their position when the event is active, and further recognizinig the sound event class.

1.2 Objectives of the Thesis

The main objective of this thesis is to develop a data-driven approach using deep neural networks (DNNs) for the SELDT task. This includes first defining the SELDT task requirements, formulating the task as a data-driven approach, and finally identifying DNNs that are best suited for the task. We define the key requirements of the SELDT task as follows. A SELDT method should be able to recognize a selected subset of sound event classes among the innumerable potential sound classes. The method should be able to detect multiple overlapping sound event trajectories in complete 2D spherical space, when the respective sound event is active.

On defining the SELDT task requirements, we identified the sub-tasks of SED and MSET from the literature that were each partially fulfilling the SELDT requirements. Together they were fulfilling our complete SELDT requirements. In this regard, we propose to perform SED and MSET jointly to produce SELDT. The two sub-tasks are each studied individually, building on the existing state of DNN approaches such as recurrent neural networks (RNNs) and convolutional recurrent neural networks (CRNNs). We propose novel sound representations and corresponding DNN adaptations to support these representations. In the process, we aim to develop an understanding of the working principles of DNNs for SED, MSET and SELDT tasks.

The main research questions studied in this thesis are as follows:

1. Can multichannel audio features help recognize overlapping sound events better than single-channel audio features for the SED task?

2. Can DNN-based DOA estimation methods estimate the number of overlapping sound events directly from input data, and localize their respective spatial locations?

3. Can SED and DOA estimation be performed jointly to produce SELDT results using DNNs?

1.3 Main Results of the Thesis

The main results and contributions of the publications leading to this thesis are as follows.

Sound Event Detection in Multichannel Audio Using Spatial And Harmonic Features

In publication [I], the first multichannel SED method exploiting the spatial and harmonic features is proposed to improve the recognition of overlapping sound events. The proposed features are inspired by the interaural intensity and time differences, and perceptual features that human auditory system use for recognizing sound events successfully in complex acoustic scenarios. An RNN with long-short term memory units is employed to

map these features to their respective sound classes. This method provides the state-of-the-art results for the SED task and won the IEEE AASP challenge on SED at DCASE 2016[1]. More about the publication is discussed in Section 4.2.

Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network

In publication [II], a CRNN-based architecture is proposed for multichannel SED that supports multiple feature classes and scales seamlessly to any number of input channels. The CRNN is evaluated on 15-times larger dataset than the one used in [I] to conclusively prove that multichannel audio helps recognize overlapping sound events better than single-channel audio. Additionally, the CRNN is also shown to learn relevant information about handcrafted features from simple low-level features. This method is the current state-of-the-art for SED and won the IEEE AASP challenge on SED at DCASE 2017[2]. More about the publication is discussed in Section 4.3.

Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features

In publication [III], the performance of overlapping sound events recognition with different spatial sampling using microphone arrays is studied. Specifically, single-channel, binaural and four-channel audio recordings of an identical acoustic scene are used with the CRNN method from [II]. It is observed that overlapping sound events are recognized better with higher spatial sampling (four-channel) than using single-channel audio features. Further, the performance of the CRNN is studied on three different acoustic scenes with up to one, three and six temporally overlapping sound events. Here it was observed that acoustic scenes with a higher number of overlapping sound events required larger CRNN. More about the publication is discussed in Section 4.4.

Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network

In publication [IV], a CRNN-based architecture is proposed for estimating the DOAs of multiple temporally overlapping and spatially stationary sound events. This is the first DNN-based method to estimate DOA in complete azimuth and horizontal angles and produce the spatial pseudo-spectrum jointly. Unlike the parametric DOA estimators that require an estimate of the number of active sources to estimate their respective DOAs, the proposed method learns the number of active sources information directly from the input features. Further, the proposed method was shown to be more robust to reverberation than parametric methods and generalize to unseen mismatched reverberant scenes. More about the publication is discussed in Section 5.2.

Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Network

In publication [V], the two sub-tasks SED and MSET were tied to produce SELDT estimates using a CRNN. Specifically, the confidence of SED from the CRNN in [II] is used to estimate a DOA trajectory for each sound class. This is the first DNN-based

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-sound-event-detection-in-real-life-audio
[2] http://dcase.community/challenge2017/task-sound-event-detection-in-real-life-audio-results

SELDT method performing localization and tracking. The proposed method is evaluated on SELDT datasets with temporally overlapping and spatially stationary sound sources. The proposed method is observed to work robustly on unseen mismatched reverberant scenes, and estimate DOA values not seen in the dataset. The method is shown to be a generic approach to jointly learn SED and MSET by evaluating on different microphone array structures. Further, to foster SELDT research using data-driven methods, the work from this publication is formalized into a task in the DCASE 2019 challenge[3]. More about the publication is discussed in Section 5.3.

Localization, Detection, and Tracking of Multiple Moving Sources with Convolutional Recurrent Neural Network

In publication [VI], the SELDT method proposed in [V] is evaluated on datasets with both spatially stationary and moving sources with varying angular velocities. It is shown that the recurrent layers employed in the SELDT method operate similarly as parametric trackers and hence enable the SELDT method to track moving sources. The MSET performance of the SELDT method is compared with a parametric MSET method. It is observed that the SELDT method recognizes and tracks multiple overlapping and moving sound events successfully, and further gives competitive MSET results in comparison to the baseline parametric method. More about the publication is discussed in Section 5.4.

1.4 Outline and Structure of the Thesis

The remainder of this thesis is organized as follows.

Chapter 2 first formulates the SED, MSET, and SELDT tasks. Their individual applications and challenges are discussed thereafter. Finally, the datasets and metrics used for evaluating these tasks are presented.

Chapter 3 presents the basic background information on acoustic feature extraction and DNNs for audio content analysis.

Chapter 4 presents approaches for SED using multichannel audio. Different spatial and perceptual features are investigated in the context of SED. Different neural network architectures are explored for the SED task to support these multiple feature classes. The proposed features and architectures are evaluated on suitable datasets.

Chapter 5 first explores the estimation of multiple DOAs of temporally overlapping and spatially stationary sound events using a CRNN. Thereafter, a SELDT approach tying the SED and MSET is proposed. The proposed SELDT approach is evaluated on datasets recorded using different microphone arrays and different acoustic scenarios. The performance of the SELDT method is compared with competing stand-alone baselines.

Chapter 6 presents the conclusions of this thesis and further discusses the future directions for SELDT research.

[3] http://dcase.community/challenge2019/task-sound-event-localization-and-detection

2 Problem Description

In this chapter, we first formulate the SELDT task. Thereafter we discuss the applications and challenges of the SELDT task. Finally, we present the datasets and metrics required for the evaluation of SELDT methods.

2.1 Problem Formulation

A real-life acoustic environment can be described either broadly by the scene, or in a finer resolution by describing the sound events it is comprised of. For example, the acoustic environment of a *park* scene, can contain sound events such as *dog barking*, *children talking*, and *birds singing*. These sound events are produced by their respective sources, for example, the dog is the source of the sound event *dog barking*. Further, each individual source is capable of producing different kinds of sound events. If child is the source, then it can produce sound events such as *laughing*, *crying*, or *talking*.

Based on the number of sound event classes to be recognized and the type of annotation for these recognized classes, there are three broad tasks - classification, tagging, and detection. Sound event classification [3, 4] is the task of recognizing a test audio sample as part of a single sound event class among many classes. Recognizing multiple sound event classes in the same test audio sample is referred to as sound event tagging [5]. Finally, estimating the temporal activities for each of the sound event classes is referred to as SED.

Localization is the task of estimating the relative position of the sound event, most often with respect to the microphone used for recording the sound event. In the context of this thesis, localization is performed by estimating the DOA of the sound event and is represented in the 2D spherical coordinate space of azimuth $\in [-\pi, \pi]$ and elevation $\in [-\pi/2, \pi/2]$ angles. In a real-life acoustic scene the sound sources producing the sound events are often not stationary but can be moving, hence the sound event moves too. The task of localizing the individual sound event's spatial trajectory when active is referred to as sound event tracking. Further, tracking multiple temporally and spatially overlapping sound events is referred to as MSET.

In this thesis, the input to the SELDT method is a multichannel audio recording of an acoustic scene recorded using a microphone array. The goal of the SELDT method is then to analyze the multichannel audio to detect the sound events of interest, and their spatial trajectories. The SELDT task can be formulated as the joint SED and MSET sub-tasks. In the SED sub-task, we only detect the C number of classes that are of our interest and are known beforehand. For the MSET sub-task, we track the spatial trajectories of the detected sound events.

2.2 Applications

The SELDT task, and its sub-tasks SED and MSET, each have numerous applications. Successful implementation of SELDT will provide context awareness to machines and will enable them to interact with the world seamlessly. Robots in a complex acoustic scenario can recognize the sound event of interest and navigate in its direction [6–9]. Smart teleconferencing hardware can recognize the active speaker and track their motion over time [10–14]. Further, the tracked location of the active speaker can be used for enhancing the speech with beamforming methods for improving automatic speech recognition (ASR). According to the World Health Organization [15], 5 % of the world's population suffer from hearing disability. With the help of SELDT, we can build assistants that will help these hearing-impaired people to visualize sounds and enable them to interact with the world naturally.

One of the important applications of SED is in the acoustic monitoring domain. In order to monitor wildlife and other biodiversity, acoustic sensors are being deployed across the forest that record audio. These sensors are small in size, easily affordable and non-invasive. The recordings from these sensors are later collected and processed to extract ecological data, including species occupancy and abundance, population density, and biodiversity [16–19]. Similarly, sensor arrays are also used to monitor urban sounds [20–22]. SONYC[1] is one such effort where across New York city acoustic sensors have been deployed to collect the noise and sound composition information. These sensors are currently monitoring the ten most common sound events in New York, such as car horn, drilling, siren, and street music, and drawing interesting correlations on the distribution of different sounds across the city. This information is being used by city experts and agencies in decision making. More about SONYC can be read in [20].

In the current era of big data, with users uploading hours of multimedia content each second of the day on the internet poses several challenges. Particularly in the context of audio content, recognizing the sound events in them will help in organizing the big data into categories, and further help in quick retrieval for user queries [23]. Another critical task with big data is to screen them for malicious or threatening content. As an example, an audio clip with a gunshot and screaming sounds suggest violent content. This can be automatically detected using a SED method and the flagged content can be blocked for consumers, including children, who opt out of such violent content. Alternatively, the same information can also be used to automatically identify the genre of movies.

Smart homes, and smart cities can employ SED or SELDT for acoustic scene analysis and surveillance [24–27]. Particularly, smart home devices with inbuilt microphone arrays such as assistants, speakers, fire and burglar alarms, are being used for detecting sound events such as glass breaking, gunshot, and smoke alarms [28, 29]. In fact, using audio for this type of scene analysis is more advantageous than using video. The detection of objects with video input is poor in low light, and when the objects are out of sight. In contrast, audio works irrespective of light, or sight and hence effectively covers a larger area than video. Further, processing and storage of audio are significantly cheaper than video and the algorithms can be implemented on low processing power chips [28].

[1] https://wp.nyu.edu/sonyc/

2.3 Challenges

Implementing a SELDT system with real-life audio is challenging in multiple aspects. One of the biggest challenges is the dataset collection. In order to train a supervised method (discussed later in Section 3.2) for a SELDT task, it needs detailed annotation of the onset-offset times, the spatial location with respect to time and the sound class. Annotating such a dataset is time-consuming and expensive. Apart from the dataset, a few of the other important challenges are intra-class variability, definitive ambiguity, overlapping sound events, spatial resolution, and recording conditions. Most of these challenges are not specific to SED, MSET or SELDT but also apply to other audio tasks such as ASR and music transcription.

Intra-class Variability

The sound classes used for SED task are often broadly defined, such as *car horn*, or *phone ring*. But these classes have a lot of variability within the class. For examples, not all cars have identical horns, even within similar model cars, there are minor variations in the horn sounds. Further, the duration of the horns and the repeated structure in which they are played is under the control of the driver. A robust SED method will have to handle such variations and predict all these different horn sounds to a single *car horn* class.

Definitive Ambiguity

Different kinds of definitive ambiguity are faced during manual annotation of audio for the SED task. The most common among them is the decision of marking the onset and offset times. Sound events such as *phone ring* or *car horn* have a distinct start and end time that are easily distinguishable. However, sound events such as *vehicle passing by* have relatively long rise and fall times, and labeling the onset and offset is often subjective to the annotator. A similar ambiguity occurs for repeating sound classes such as *foot steps* or *hammer* that most often occur multiple times within close intervals. Although a powerful classifier can potentially recognize each instance of the *foot step* and *hammer*, the annotator might label all the instances together as one event. Similar ambiguity also appears for sound events in a low SNR scenario. Here the annotator will have to make a subjective decision about whether the sound event is audible enough to be recognized by a classifier or to be marked as background noise. Finally, for real-life recordings, annotating some isolated sound events can be ambiguous without context or visuals of the scene. This ambiguity in recognizing sounds has been traditionally used to create sound effects for films using props, and is referred as foley[2,3]. As an example, the sound of thunder can be created by shaking a thin metal sheet. Without the visuals, this sound from the metal sheet will be perceived as thunder.

Overlapping Sound Events

The sound events in real-life acoustic scenes are polyphonic in nature, i.e., multiple sound events are temporally overlapping with each other. The resulting audio recording will have a mixture of these overlapping sound events. Further, the mixtures keep changing over time based on the different overlapping sound events. A robust SED method should be able to recognize all the overlapping sound events in the different mixtures.

[2]https://en.wikipedia.org/wiki/Foley_(filmmaking)
[3]http://www.marblehead.net/foley/whatisitman.html

These overlapping sound events are also challenging for localization. In fact, some maximum likelihood-based localization methods only support localization of $K-1$ overlapping sound events given a K-channel recording [30]. Recent methods are proposing to overcome this restriction and estimate more sources than the number of microphones in the array [31–34]. Most parametric localization methods require information on the number of active sources in order to estimate their respective DOAs. As this information is not always available, counting schemes such as minimum description length [35] and Akaike information criterion [36] have been employed. However, the performance of these schemes in a real-life dynamic environment is poor. Therefore, more recent methods have been proposing to learn the number of sources directly from the input features using data-driven DNN-based methods [IV-VI][9, 37].

Spatial Resolution

The resolution of the localization method becomes more critical for detection of spatially close sources. Most parametric MSET methods produce localization estimates at a fixed spatial resolution. One of the ways to improve localization performance is to increase the spatial resolution. However, this results in increased search space (azimuth and elevation) for estimating the DOA and hence longer processing times. Similarly, for the DNN-based methods using a supervised classification approach (see Section 3.2), higher resolution implies a large number of output classes. This number of classes increases further for classification-based SELDT approach, where the number of sound classes is multiplied with the number of spatial locations. Training such DNNs with a large number of classes, where the number of positive classes is few and the negative classes are large, results in imbalanced dataset problems and the challenges associated with it. Further, the dataset size grows rapidly with the number of classes, as a result of collecting sufficient samples for each class.

Recording Conditions

In the scope of the SED task, other than the pre-defined set of target sound event classes the remaining sound events are considered background noise. Some of these background noises, for example, wind can be loud, making the signal to noise ratio (SNR) of the sound events of interest low. Recognizing and further localizing such low SNR sound events can be challenging. Further, the type of recording hardware used, and whether it is professional or consumer grade, may present challenges to the SED methods. One such challenge can arise from the mismatch in microphone frequency responses, DNN models trained on a particular microphone response may not perform consistently across other microphones with different frequency responses.

Reverberant scenarios that commonly occur in real-life can be challenging for the localization task [38]. Furthermore, the performance of localization methods is known to vary with changes in temperature, humidity, and atmospheric pressure [39]. This is a result of the change in speed of sound through the air which is required for modeling the localization problem. Additionally, the information of the microphone array geometry configuration and its respective microphone properties are crucial for successful localization. However, the array geometry, i.e., the spacing between microphones can be variable, e.g., in robot audition [40]. An off-the-shelf localization method should be able to handle and function independently of these hardware changes.

2.4 Evaluation

To evaluate the SED, MSET and SELDT methods, we employ datasets that sample the different acoustic scenarios to be studied and define a set of metrics to quantitatively measure their performance in this section.

2.4.1 Datasets

The datasets used for evaluation in this thesis are presented in Table 2.1. Among the twelve datasets, apart from the TUT-SED 2009, the remaining eleven datasets are publicly available enabling reproducibility of the proposed research methods. The datasets vary in terms of the hardware configuration used for the recording – binaural, four-channel first-order Ambisonic (FOA) and eight-channel circular arrays. The FOA recordings are a popular spatial audio format to record 360° audio. The four-channels of FOA are commonly referred as the W, X, Y and Z, where the X, Y and Z channels represent the pressure gradient along the corresponding x, y and z axes of the Cartesian coordinate system, and the W channel corresponds to omnidirectional pressure. With respect to the spatial response of the microphones used, the binaural format and the X, Y, Z channels of the Ambisonic format have directional responses. Whereas the microphones in the circular array and the W channel of Ambisonic format have omnidirectional responses. Unlike the directional microphones which encode the directional information predominantly in the magnitude difference, the omnidirectional microphone arrays encode this information in both magnitude and phase differences.

The datasets studied in this thesis can be broadly grouped into anechoic and reverberant acoustic scenes. Particularly the SED datasets have recordings from thirteen reverberant acoustic scenes such as home, park, street, and residential area enabling the study of performances in these individual scenes. All the SED-only datasets studied are real-life recordings, whereas the SELDT datasets are all synthesized using simulated or real-life impulse responses (IRs). In general, the task of annotating real-life recordings for SED is a time consuming and expensive task. To overcome this, most recent works are exploring the learning of SED from datasets that provide only the sound event labels without any temporal information [41 – 47]. In the case of SELDT tasks, the spatial locations for individual sound events with respect to time have to be annotated in addition to SED annotation. This makes the SELDT annotation task significantly more complex than the SED annotation. Hence in this thesis, we only study the SELDT performance with realistic synthesized datasets and consider collecting and annotating real-life SELDT datasets for future studies. Additionally, using the synthesized datasets provides accurate time-boundaries and spatial location trajectories and enables a fair evaluation of the proposed methods.

In order to have real-life complexities in the synthesized datasets, we employ measured IR to spatially position sound events in some of our datasets. Further, during synthesis, the isolated sound events used are sampled from a large dataset and comprise of inter- and intra-class variability such as a different number of classes, and variety of examples within each class. The description of individual datasets is provided below.

TUT-SED 2009

The TUT-SED 2009 dataset [48] consists of 103 binaural recordings from ten different acoustic scenes. Among the ten scenes, three are outdoor (beach, street, track and field)

Table 2.1: Datasets used for evaluation in this thesis

Dataset	Format (# channels)	Acoustic scene	Total length (in mins)	# classes	Sound location annotation	Impulse response	Acronym
SED only							
TUT-SED 2009	Binaural (2)	Reverberant	1133	61	×		
TUT-SED 2016 Development	Binaural (2)	Reverberant	78	18	×		
TUT-SED 2017 Development	Binaural (2)	Reverberant	70	6	×		
SELDT - Static sources							
TUT Sound Events 2018	Ambisonic (4)	Anechoic	450	11	✓	Synthetic	ANSYN
TUT Sound Events 2018	Ambisonic (4)	Reverberant	450	11	✓	Synthetic	RESYN
TUT Sound Events 2018	Ambisonic (4)	Reverberant	450	8	✓	Real-life	REAL
TUT Sound Events 2018	Ambisonic (4)	Reverberant	1125	8	✓	Real-life	REALBIG
TUT Sound Events 2018	Ambisonic (4)	Reverberant	1125	8	✓	Real-life	REALBIGAMB
TUT Sound Events 2018	Circular (8)	Anechoic	450	11	✓	Synthetic	CANSYN
TUT Sound Events 2018	Circular (8)	Reverberant	450	11	✓	Synthetic	CRESYN
SELDT - Moving sources							
TAU Moving Sound Events 2019	Ambisonic (4)	Anechoic	450	11	✓	Synthetic	MANSYN
TAU Moving Sound Events 2019	Ambisonic (4)	Reverberant	450	8	✓	Real-life	MREAL

and seven are indoor (basketball court, bus, car, hallway, office, restaurant, and shop). Each of these scenes consists of 8-14 recordings (10-30 minutes) and amounts to 1133 minutes in total. These recordings have been manually annotated into 61 sound event classes with an average polyphony of 2.53. Each scene has about 9-16 classes, with some scene-specific classes and the rest occurring across scenes. The annotation was done by two semi-professionals (post-graduate students), with no overlapping recordings between the two. The dataset provides predefined training, validation and testing splits.

TUT-SED 2016 Development

The TUT-SED 2016 Development dataset [49] consists of 22 binaural recordings from two scenes – home and residential area. Each of the 22 recordings is 3-5 minutes long amounting to 78 minutes in total. These recordings have been manually annotated into 11 sound event classes in the home scene and seven in the residential area. The TUT-SED 2016 dataset is available publicly[4]. The dataset provides predefined four-fold training and testing splits.

TUT-SED 2017 Development

The TUT-SED 2017 Development dataset [50] consists of 24 binaural recordings from the street scene. Each of the recordings is 3-5 minute long amounting to 70 minutes in total. These recordings have been manually annotated into six sound event classes. The TUT-SED 2017 dataset is available publicly[5]. The dataset provides predefined four-fold training and testing splits.

ANSYN

The TUT Sound Events 2018 - Ambisonic, Anechoic and Synthetic IR dataset (ANSYN) [III, IV, V] consists of sound events that are spatially positioned using synthetic IRs simulated in an anechoic environment. The dataset provides three subsets with a different number of overlapping sound events: no temporally overlapping sources ($O1$), maximum two temporally overlapping sources ($O2$) and maximum three temporally overlapping sources ($O3$). Each of these subsets has predefined three fold splits of 240 training recordings and 60 testing recordings of length 30 s, together amounting to 450 minutes of data for an individual subset. The dataset is synthesized with the 11 classes of isolated sound events from the DCASE 2016 task 2 dataset [51] such as speech, coughing, door slam, page-turning, phone ringing, and keyboard. The DCASE 2016 dataset provides 20 examples for each of these 11 classes, which are randomly split into 80% training and 20% testing for each fold. In order to synthesize a recording for the training split, for example, the $O2$ subset training split, isolated sound examples are randomly chosen from the training subset of examples and are each associated with a temporal start time such that across the recording a maximum of only two sound events overlaps temporally. Further, each of the sound examples is randomly placed in a spatial grid of complete azimuth angle $\in [-180°, 180°)$ and elevation angle $\in [-60°, 60°)$, both at 10° resolution. Additionally, in order to have amplitude variability across instances of the same sound example, each example is randomly placed at a distance of $\in [1, 10]$ m with 0.5 m resolution from the microphone. A similar procedure is carried out for other subsets and their respective cross-validation splits. The ANSYN dataset is available publicly[6]. More details on the

[4]https://zenodo.org/record/45759
[5]https://zenodo.org/record/814831
[6]https://zenodo.org/record/1237703

synthesis of the recordings can be read in [III, IV, V].

RESYN

The TUT Sound Events 2018 - Ambisonic, Reverberant and Synthetic IR dataset (RESYN) [IV, V] is synthesized identically to ANSYN with all the sound events spatially located within a reverberant room. During simulation, the microphone is placed in the center of the room of dimension $10 \times 8 \times 4$ m (Room 1), with reverberation times 1.0, 0.8, 0.7, 0.6, 0.5, and 0.4 s per each octave band, and 125 Hz–4 kHz band center frequencies. The isolated sound event examples from the DCASE 2016 task 2 dataset [51] are spatially placed around the microphone in complete azimuth angle $\in [-180°, 180°)$ and elevation angle $\in [-60°, 60°)$ both at 10° resolution, and at a distance ranging from 1 m to the respective end of the room at 0.5 m resolution. Similar to ANSYN, RESYN has three subsets $O1$, $O2$, and $O3$ with a different number of overlapping sound events. Further, each of the subsets has three cross-validation training and testing splits. Additionally, the RESYN datasets provide testing splits for two different sized rooms: Room 2 that is 80% the volume ($10 \times 8 \times 4$ m) and reverberation time of Room 1. Room 3 that is 125% the volume ($10 \times 10 \times 4$ m) and reverberation time of Room 1. The testing splits of Room 1, 2 and 3 are identical in terms of the temporal and spatial distribution of the sound events. Individual sound events whose distance from the microphone exceeded the room size were assigned a new distance to fit the new room dimension. Unless the room information is specifically mentioned, the RESYN dataset is synonymous to the Room 1 dataset. The RESYN recordings are available publicly[7]. More details on the synthesis of the recordings can be read in [IV, V].

REAL

The TUT Sound Events 2018 - Ambisonic, Reverberant and Real-life IR dataset (REAL) [V] is synthesized in a similar fashion as ANSYN and RESYN but with real, measured IRs collected using an Eigenmike[8] spherical microphone array. In order to collect the IR, a maximum length sequence (MLS) was played back using a Genelec G Two[9] loudspeaker while moving slowly in a circular trajectory around the Eigenmike as shown in Figure 2.1. Finally, the moving-source IR is obtained from the MLS recorded by the Eigenmike using the publicly available tool from the CHiME challenge [52]. This tool estimates the time-varying IR in the short-time Fourier transform domain by assuming block-wise stationarity of the acoustic signal. The IR is obtained by solving the least-squares regression between the known measurement signal and the far-field recording independently at each frequency. The IRs were recorded in a university corridor with classrooms around it. During the recording, the average playback volume was set to be 30 dB greater than the ambient sound level. The recorded IRs are available publicly[10].

For the dataset creation, we collected IRs at elevations in the range $\in [-40°, 40°]$ at 10° resolution at 1 m distance from the Eigenmike, and $\in [-20°, 20°]$ at 10° resolution at 2 m distance. Further at each of these elevation-distance pairs, IRs were collected in the complete azimuth at 10° resolution, i.e., 36 IRs for each elevation-distance pair. In order to synthesize the dataset we used the real-life isolated sound events from the urbansound8k [3] dataset. This dataset contained 10 classes such as air conditioner, car

[7] https://zenodo.org/record/1237707
[8] https://mhacoustics.com/products
[9] https://www.genelec.com/home-speakers/g-series-active-speakers
[10] https://zenodo.org/record/1443539

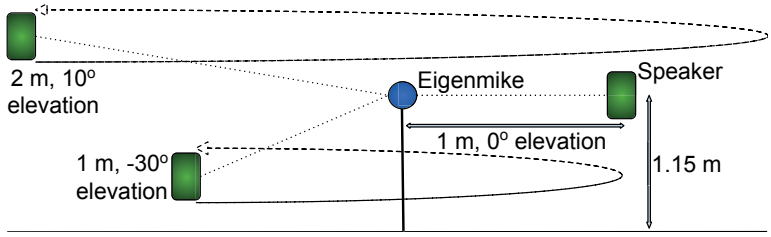


Figure 2.1: A schematic diagram for real-life impulse response collection. A speaker playing maximum length sequence is slowly moved around the Eigenmike to complete a circle at different distances and elevations from the Eigenmike. ©2018 IEEE.

horn, children playing, dog barking, drilling, engine idling, gunshot, jackhammer, siren and street music. Among these, we discarded the children playing and air conditioner classes since these could also occur in the acoustic scene of an university corridor, which we use as background in our REALBIGAMB dataset (discussed later). Finally, to synthesize a recording we choose an isolated sound event example randomly and convolve it with a randomly chosen IR to spatially position it, and further randomly assign a temporal location within a 30 s long recording. Additionally, similarly to the previous datasets, three subsets $O1$, $O2$ and $O3$ with a different number of temporally overlapping sound events are synthesized, each with three-fold cross-validation splits having 240 training and 60 testing recordings. The synthesized dataset is available publicly[11]. More details on the synthesis of the recordings can be read in [V].

REALBIG

The TUT Sound Events 2018 - Ambisonic, Reverberant and Real-life IR big dataset (REALBIG) [V] is an exact replica of the REAL dataset in terms of synthesis but contains 2.5-times more recordings than the REAL dataset. The cross-validation split for each of the three subsets ($O1$, $O2$, and $O3$) contains 600 recordings for training and 150 for testing. This dataset is larger than the open source hosting website budget, hence available on request.

REALBIGAMB

The TUT Sound Events 2018 - Ambisonic, Reverberant, Real-life IR and Ambiance big dataset (REALBIGAMB) [V] is an exact replica of REALBIG dataset, but additionally contains background ambiance. A 30 min long ambiance recording was collected in the same location as the IR recordings, without changing the setup. Randomly chosen ambiance segments were then added to the recordings from the REALBIG dataset at three different signal to noise ratios (SNRs): 0, 10 and 20 dB for each of the three subsets ($O1$, $O2$, and $O3$). This dataset is larger than the open source hosting website budget, hence available on request.

CANSYN

The TUT Sound Events 2018 - Circular array, Anechoic and Synthetic IR dataset (CANSYN) [V] is an exact replica of the ANSYN dataset but synthesized using circular

[11] https://zenodo.org/record/1237793

array IRs. The circular array used has 5 cm radius with eight omnidirectional microphones at a 45° interval, and the array plane parallel to the ground. Similar to previous datasets, CANSYN has three subsets $O1$, $O2$, and $O3$ with a different number of overlapping sound events. The dataset is publicly available[12].

CRESYN

The TUT Sound Events 2018 - Circular array, Reverberant and Synthetic IR dataset (CRESYN) [V] is an exact replica of RESYN dataset but synthesized using circular array IRs. The circular array used is similar to CANSYN and placed in the center of the room with the array plane parallel to the floor during synthesis. Similar to the previous datasets, it consists of three subsets $O1$, $O2$, and $O3$ with a different number of overlapping sound events. The dataset is publicly available[13].

MANSYN

The TAU Moving Sound Events 2019 - Ambisonic, Anechoic, Synthetic IR and Moving Source dataset (MANSYN) [VI] is identical to ANSYN in terms of the acoustic environment and sound events. However, instead of having spatially stationary sound sources, they are simulated to have a constant angular velocity in the range $\in [-90°, 90°]/s$ with $10°/s$ steps. Further, the resolution of the spatial grid used to generate the IR is reduced to 1° from 10° in ANSYN. The sound events in this dataset are moving along both the azimuth and elevation. More about the dataset synthesis can be read in [VI]. The dataset is publicly available[14].

MREAL

The TAU Moving Sound Events 2019 - Ambisonic, Reverberant, Real-life IR and Moving Source dataset (MREAL) [VI] is identical to REAL in terms of the acoustic environment and sound events. However, the sound events are synthesized to be moving around with a constant angular velocity in the range $\in [-90°, 90°]/s$ with $10°/s$ steps. The measured IR recordings from REAL are employed for MREAL, but for this dataset, the IRs were sampled at 1° resolution in azimuth. Since these IR were collected only for motion along azimuth, the synthesized sound sources in this dataset also have motion only along azimuth. More about the dataset synthesis can be read in [VI]. The dataset is publicly available[15].

2.4.2 Metrics

The SED and DOA trajectory estimation performance of the methods proposed in this thesis are evaluated with separate metrics.

SED metrics

As the SED metrics, we use the standard polyphonic metrics, error rate (ER) and F-score calculated in one-second segments with no overlap as proposed in [53, 54]. These metrics were also used as the official metrics for the SED tasks in the IEEE AASP challenges

[12] https://zenodo.org/record/1237752
[13] https://zenodo.org/record/1237754
[14] https://zenodo.org/record/2636586
[15] https://zenodo.org/record/2636594

DCASE 2016 [55] and 2017 [50]. A sound event is said to be active in a one-second segment if it is active in at least one of the time frames (40 ms) within the segment. The segment-wise F-score is then defined as

$$F = \frac{2 \cdot \sum_{k=1}^{K} TP(k)}{2 \cdot \sum_{k=1}^{K} TP(k) + \sum_{k=1}^{K} FP(k) + \sum_{k=1}^{K} FN(k)}, \tag{2.1}$$

where the number of true positives $TP(k)$ is the total number of sound event classes that were active in both reference and predicted one-second segment k. Similarly, the number of false positives $FP(k)$ is the total number of sound event classes that were active in the predictions but inactive in the reference, and the number of false negatives $FN(k)$ is the total number of sound event classes that were active in the reference but inactive in the prediction.

The ER is defined as

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)}, \tag{2.2}$$

where $N(k)$ is the total number of active sound event classes in the reference one-second segment k. The intermediate statistics, substitutions $S(k)$, deletions $D(k)$ and insertions $I(k)$ are defined using just the $FN(k)$ and $FP(k)$ metrics for the SED task as:

$$S(k) = \min(FN(k), FP(k)), \tag{2.3}$$

$$D(k) = \max(0, FN(k) - FP(k)), \tag{2.4}$$

$$I(k) = \max(0, FP(k) - FN(k)). \tag{2.5}$$

Substitution is obtained by merging the false negatives and positives without assigning which false positives substitute which false negatives. The remaining false negatives are counted as deletions and the remaining false positives are counted as insertions.

An SED method is evaluated jointly using the F-score and ER. An ideal SED method will produce an ER of zero and F-score of one.

MSET metrics

An individual DOA is represented by its azimuth ϕ and elevation θ angles as (ϕ, θ). The spatial pseudo-spectrum (SPS) represented as $SPS(\phi, \theta)$ is then defined as the intensity of the sound source in the direction (ϕ, θ). The SPS estimated by the proposed method $SPS_E(\phi, \theta)$ is evaluated with respect to a baseline $SPS_{GT}(\phi, \theta)$ using the signal to noise ratio metric defined as:

$$SNR = 10 \log_{10} \Big(\frac{\sum_{\phi} \sum_{\theta} SPS_{GT}(\phi, \theta)^2}{\sum_{\phi} \sum_{\theta} (SPS_E(\phi, \theta) - SPS_{GT}(\phi, \theta))^2} \Big). \tag{2.6}$$

An MSET method generates multiple DOA trajectories for each input audio recording. A DOA trajectory is defined as a list of DOAs from consecutive time-frames, corresponding to a single sound event instance. In this thesis, since we do not assume any limit to the number of temporally overlapping sources, each time-frame can have multiple DOAs corresponding to different DOA trajectories.

The MSET performance of the proposed methods is evaluated using two frame-wise metrics, the DOA error and frame recall. The DOA error is the average angular error in

degrees between the estimated and reference DOAs. For a recording of length T time frames, let D_R^t represent the number of reference DOAs in the t-th time frame and D_E^t the number of estimated DOAs. The DOA error is then defined as

$$DOA\,error = \frac{1}{\sum_{t=1}^{T} D_E^t} \sum_{t=1}^{T} \mathcal{H}(\mathbf{DOA}_R^t, \mathbf{DOA}_E^t), \tag{2.7}$$

where $\mathbf{DOA}_R^t$ and $\mathbf{DOA}_E^t$ are the list of reference and estimated DOAs at time frame t. $\mathcal{H}()$ is the Hungarian algorithm [56] for solving the assignment problem, i.e., matching the individual estimated DOAs in a time-frame with the respective reference DOAs. The Hungarian algorithm solves this assignment by minimizing the pair-wise cost between the individual reference and estimated DOA. As the pair-wise cost to minimize we use the central angle σ which is the angle between the estimated and predicted angle at the center of the sphere

$$\sigma = \arccos(\sin\theta_E \sin\theta_R + \cos\theta_E \cos\theta_R \cos(\phi_R - \phi_E)) \tag{2.8}$$

where the reference DOA, is represented by the azimuth angle $\phi_R \in [-\pi, \pi)$ and elevation angle $\theta_R \in [-\pi/2, \pi/2]$ and the estimated DOA is represented with (ϕ_E, θ_E).

In order to account for time-frames where the number of estimated and reference DOAs are unequal, we report the second metric frame recall, which is calculated as,

$$Frame\,recall = \frac{\sum_{t=1}^{T} \mathbb{1}(D_R^t = D_E^t)}{T}, \tag{2.9}$$

where $\mathbb{1}()$ is the indicator function resulting in an output one if the $(D_R^t = D_E^t)$ condition is met else returns zero.

In this thesis, the DOA error is reported in degrees, and an ideal MSET method results in a frame recall of one and DOA error of zero.

Early stopping metrics

Since there are two metrics each for both SED and MSET, we can choose one among the two metrics for early stopping (discussed in Section 3.4) for the respective method. Alternatively, we can use a weighted combination of the metrics for early stopping. In this regard, the SED training can be stopped with a single metric, SED score, defined as

$$SED\,score = (ER + (1 - F))/2. \tag{2.10}$$

The two MSET metrics can together be represented by a single metric, the MSET score, defined as

$$MSET\,score = \big(DOA\,error/180 + (1 - frame\,recall)\big)/2, \tag{2.11}$$

and the joint SED and MSET performance can be measured using

$$SELDT\,score = (SED\,score + MSET\,score)/2. \tag{2.12}$$

An ideal method will have the SED, MSET and SELDT score of zero.

3 Background

In this chapter, we present a brief introduction to audio content analysis using machine learning methods, and further relate them to the SELDT task. Most of the audio content analysis methods have two main stages, the sound representation, and the classification. More about the two stages are discussed in this chapter.

3.1 Sound Representation

The audio recordings for the SELDT task are recorded digitally. This produces a time-domain representation that is the lowest form of acoustic features. In comparison, the human auditory system has evolved to analyze and understand the acoustic scene around it using a frequency-domain representation. This representation has shown to be more abstract than the time-domain, and hence less redundant in the information it contains and more robust to noise. For this reason, most audio content analysis methods use some form of frequency-domain representation.

Spectrogram

The most common frequency-domain representation of audio is obtained using an algorithm referred to as the short-time Fourier transform (STFT). This assumes the audio is stationary in short time-frames and is a sum of sinusoids of different frequency. Given a time-domain audio signal, the output of STFT provides the composition of different frequency sinusoids in each of the short time-frames resulting in the frequency-domain representation, referred to as spectrogram hereafter. The short time-frames are generally obtained by using a windowing function of length 20 to 50 ms with smooth tapering on either end. This window is multiplied with the time-domain representation of the audio in hop-lengths of 25 % to 50 % of the window length. At each hop-length, the product of the window and the audio is transformed to the frequency domain using a discrete Fourier transform (DFT) algorithm such as fast Fourier transform (FFT). The operations of windowing, shifting by hop-length, and extracting the DFT together is performed by the STFT algorithm. Hamming, Hanning and Blackman functions are some examples of commonly used windows. Further, longer window length produces a higher frequency resolution of STFT. Hence the window length can be chosen based on the application.

The spectrogram is complex valued and comprises of a magnitude and phase component. Traditionally the magnitude component has been the most informative for audio content classification tasks [57], such as SEC and SED [V, VI]. Whereas the phase component is particularly informative for estimating sound source directions, and has been used for localization [IV][37, 58] and SELDT [V, VI] tasks.

Mel Spectrogram

The spectrogram has a linear frequency resolution. But empirical results in [59] showed that the human auditory system is not equally sensitive across all frequencies, rather it is more sensitive to changes in the lower frequencies than the higher frequencies. Mel scale [60] is one widely used non-linear frequency scale, that provides similar sensitivity as the human auditory system from a spectrogram. The mel scale is implemented as a mel filterbank, with multiple triangular filters distributed across the frequency range at different central frequencies. The distance between neighboring central frequencies and the corresponding frequency band of the filter at each of these central frequencies widen with the increase in frequency. Thus applying the mel filterbank on the spectrogram results in weighted averaging of magnitudes across frequencies, with higher frequency resolution at lower frequencies and lower resolution at higher frequencies. Often, in place of the spectrogram, the energy spectrogram is used with mel filterbank, resulting in the mel-band energy spectrogram. This is further processed with a logarithm operation to compress the dynamic range between the dominant energies in the low frequencies in comparison to high frequencies, referred as the log mel-band energy spectrogram or just log mel-band energies hereafter. Currently, the log mel-band energy is the most popular feature for audio content recognition algorithms, especially in SED [I-III][61 – 63]. Most often the number of filterbanks used is mel-band energy feature in the range of 40 to 80, which is significantly smaller than the number of frequency bins in the STFT. Hence, mel-band energy feature can be considered to provide more abstract and compact representation than STFT.

3.2 Machine Learning

The output of the sound representation stage is the acoustic features $\mathbf{X}$ such as phase and magnitude components of the spectrogram. These features are mapped to the SELDT output $\mathbf{Y}$ using an acoustic model represented with parameters $\mathbf{W}$. A common approach to model $\mathbf{W}$ is using machine learning, which is the science of using statistical models, without using explicit hand-crafted rules, to perform a specific task. These statistical models are learned from sample data that represents the task. In the context of this thesis, the specific task is SELDT, and sample data is commonly referred to as a dataset.

There are three major divisions in machine learning algorithms, supervised, semi-supervised and unsupervised algorithms. As the name suggests, supervised algorithms require the target labels $\mathbf{Y}$ in addition to the input acoustic features $\mathbf{X}$ for the entire dataset. On the other hand, unsupervised algorithms do not require any labels $\mathbf{Y}$, whereas semi-supervised algorithms only need the target labels $\mathbf{Y}$ for a subset of the dataset. In this thesis, we learn the acoustic model parameters $\mathbf{W}$ using supervised learning algorithms.

Based on the type of learning the machine learning algorithms can be broadly categorized into regression and classification algorithms. Regression algorithms learn a mapping between the input features and continuous valued output. On the other hand, the classification task maps the input features to a discrete set of classes.

The machine learning algorithms can also be categorized based on the output format. A two class classifier is commonly referred to as binary classifier. But if the classifier is recognizing more than two classes, and outputs only one class for each input feature, it is called multiclass classifier. On the other hand, if it outputs more than one class for each input feature then it is referred to as multiclass multilabel classification. Since in real

acoustic environments sound events are both temporally and spatially overlapping, each analysis time-frame can have more than one active sound event and hence the SELDT is a multiclass multilabel classification task.

Some of the earliest SED [64] and SELDT [14, 65] methods employed machine learning algorithms combining Gaussian mixture models (GMMs) and hidden Markov models (HMMs). Individual GMMs are learned for each sound class in the dataset, and an HMM is used to learn the temporal patterns of the respective sound classes. The GMM-HMM algorithm was first introduced for ASR and music transcription tasks and was later adapted to the SED given the task similarities of recognizing time-varying sounds and their temporal information. GMM is an example of a generative algorithm, i.e., a GMM only models the distribution of individual classes. However, it overlooks the differences and similarities with other classes in the dataset, which can potentially be used to classify better. Algorithms that use this information of neighboring classes to learn a boundary between individual classes are referred to as a discriminative algorithms. Support vector machine (SVM) is a popular discriminative algorithm that learns an optimal hyperplane separating two classes. It was shown in [66] that SVMs outperform GMMs for the SED task. Similar SVMs were also employed for SED in [67, 68].

3.3 Deep Neural Networks - Architecture

Neural networks are a class of supervised algorithms that are inspired by the human neural system and have achieved good results on a wide variety of machine learning tasks. The basic unit of the human brain, neurons, are interconnected with neighboring neurons. Each neuron receives input signals from multiple connected neurons and produces an output signal based on some inherent rules, and further, the output is forwarded to other connected neurons. Similarly, a single computational neuron can be defined as $z = f(\sum_i w_i x_i + b)$, where inputs x_i from different connected neurons are used to produce output z that is forwarded to the other connected neurons. The non-linearity function $f()$, weights $\mathbf{w}$ and bias b are parameters used to learn the inherent rules within a single computational neuron to produce output z. The main function of bias is to provide each neuron a trainable constant value in addition to the weighted input. The non-linearity function, commonly know as activation function, enables the neuron to successfully approximate non-linear decision boundaries between classes. More about the activation functions are discussed later.

Most often, for classification problems more than one neuron is used in a single layer and is defined as $\mathbf{z} = f(\mathbf{W}\mathbf{x} + \mathbf{b})$. A single input feature vector $\mathbf{x} \in \mathbb{R}^{F\times 1}$ of length F is scaled with C neurons (equal to the number of classes in the dataset) of weights $\mathbf{W} \in \mathbb{R}^{C\times F}$ and bias $\mathbf{b} \in \mathbb{R}^{C\times 1}$ and passed through a non-linearity $f()$ to obtain the class-wise results $\mathbf{z} \in \mathbb{R}^{C\times 1}$. Further, a harder classification problem might need a higher number of neurons and layers. Such networks with more than one layer are often referred to as deep neural networks (DNN). In DNNs, the first layer is often called the input layer, while the last layer is called the output or the classification layer. The remaining layers between the input and output layers are referred together as the hidden layers. For example, the output of the output layer in a three layer network is given as $\mathbf{z}^{(3)} = f(\mathbf{W}^{(3)}\mathbf{z}^{(2)} + \mathbf{b}^{(3)})$, where, $\mathbf{z}^{(2)} = f(\mathbf{W}^{(2)}\mathbf{z}^{(1)} + \mathbf{b}^{(2)})$ is the output of the hidden layer, and $\mathbf{z}^{(1)} = f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)})$, $^{(2)}$ is the output of the input layer. Since each neuron in a layer is connected with every neuron in the previous layer, this network architecture is commonly referred to as the fully-connected (FC) neural network. A similar FC network

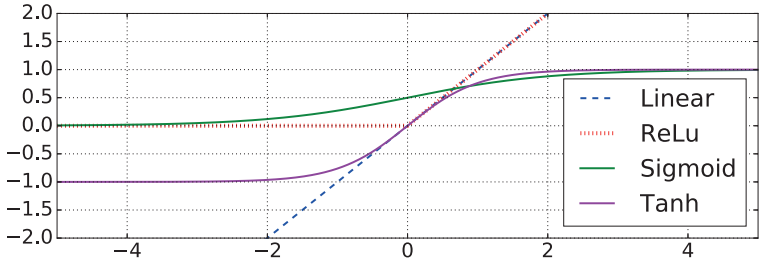


Figure 3.1: Commonly used activation functions. The vertical axis represents the output of the activation function for the corresponding input shown in the horizontal axis.

was shown to perform significantly better than GMM-HMM in [69]. This suggests that DNNs are powerful classifiers in comparison to GMM-HMM.

Activation Functions

The activation function is a crucial part of the neural network that helps it to learn complex shaped decision boundaries. Some of the activation functions used in this thesis are visualized in Figure 3.1. The vertical axis represents the output of the activation function for the corresponding inputs shown in the horizontal axis. In simple classification tasks, where the classes are easily separable with a linear decision boundary, a linear activation function $f(x) = x$ can be employed. Additionally, in the regression tasks, if the required output $\in \mathbb{R}$, a linear activation function can be employed in the output layer. For example, when approaching MSET as a regression task, we can employ linear activation in the output layer to estimate the azimuth and elevation angles.

Most real-world classification problems have complex decision boundaries and hence non-linear activation functions are required. The sigmoid activation function defined as $\sigma(x) = 1/(1 + e^{-x})$, takes real values as input and produces outputs $\in [0, 1]$. Large negative numbers are forced to zero, while large positive numbers are forced to one. Since the output range is bounded between zero and one, the sigmoid output can be considered probabilities. Hence sigmoid is mostly used in the classification layer. In the case of SELDT, sigmoid functions are used for multiclass multilabel classification required for SED.

The tanh activation function is defined as $f(x) = \tanh(x)$. The tanh function produces outputs $\in [-1, 1]$, i.e., it forces large negative values to -1, and large positive values to 1. The tanh is a shifted and scaled version of sigmoid, and their relation is given by $f(x) = 2\sigma(2x) - 1$. Since the tanh output is zero-centered in practice it is always preferred over sigmoid. In the case of SELDT, the tanh activation is used in the recurrent layers that are discussed later.

A rectified linear unit (ReLU) function is the most popular non-linear activation in the recent years, and is defined as $f(x) = \max(0, x)$. In comparison to other non-linear activation functions tanh and sigmoid, the implementation of ReLU is not expensive and hence greatly accelerates the learning process. In the case of SELDT, the ReLU activation is used in the convolutional layers that are discussed later.

Objective function

Objective functions compute the distance (positive valued) between the DNN estimate and the reference output values. Some common objective functions are the mean squared error (MSE) for regression tasks, and the binary cross entropy loss for multiclass multilabel classification tasks. The MSE is defined as $L_{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$, where $\hat{y}_n$ is the prediction and y_n is the reference for each of the N samples. The L_{MSE} produces larger values when y_n and $\hat{y}_n$ are dissimilar and smaller values when similar. In the case of regression-based approached of MSET, we use the MSE.

The binary cross entropy (CE) loss is defined as

$$L_{CE} = \frac{1}{NC}\sum_{n=1}^{N}\sum_{c=1}^{C}\Big((1 - Y_{n,c})\log(1 - \hat{Y}_{n,c}) - Y_{n,c}\log(\hat{Y}_{n,c})\Big), \tag{3.1}$$

where $Y_{n,c}$ and $\hat{Y}_{n,c}$ are the reference and prediction for the n-th sample of class c. Similar to MSE, the CE loss is larger when the $Y_{n,c}$ and $\hat{Y}_{n,c}$ are dissimilar and smaller when similar. The CE loss is used for the multiclass multilabel SED task.

3.4 Deep Neural Networks - Learning

In the previous section we briefly described the static part of DNNs, that is the architectural components. In this section we briefly present the dynamic part of the DNNs, that is the process of learning the parameters of the DNN.

Optimization Algorithms

Assume a generic neural network with parameters $\mathbf{w}$, and a sufficiently large dataset with input features $\mathbf{x}$ and target labels $\mathbf{y}$. The goal of an optimization algorithm is to find the best parameters $\mathbf{w}$ that minimize the loss $L_{\mathbf{w}}(\mathbf{y}, \hat{\mathbf{y}})$ between the predicted $\hat{\mathbf{y}}$ and target labels $\mathbf{y}$. One of the popular optimization algorithms is the gradient descent algorithm. This algorithm optimizes by iterating multiple times over the following three steps – forward propagation, loss and gradient estimation, and backward propagation. Before forward propagation, the parameters $\mathbf{w}$ are randomly initialized with small values, for example drawn from a normal distribution. During forward propagation, an input vector $\mathbf{x}$ is processed to obtain the prediction $\hat{\mathbf{y}}$ using the neural network parameters $\mathbf{w}$. Next, the quality of the prediction is computed using a corresponding loss function $L_{\mathbf{w}}$ such as the MSE or CE loss. Further, the gradient of the loss with respect to each parameter $\mathbf{w}$ given by $\frac{\partial L_{\mathbf{w}}}{\partial \mathbf{w}}$ is computed. This gradient represents the slope of the loss function $L_{\mathbf{w}}$ and is helpful in finding the direction of local loss minima. The term gradient descent implies using gradient information for descending to loss minima. Finally, during backward propagation, the network parameters are updated with the loss gradient such that the new parameters produce a lower loss $L_{\mathbf{w}}$. The network parameter update is given by $\mathbf{w} \leftarrow \mathbf{w} - \alpha\frac{\partial L_{\mathbf{w}}}{\partial \mathbf{w}}$, where the $\leftarrow$ is an assignment operator, and α is a positive number representing the learning rate which controls the magnitude of the update. This backward propagation for the neural network parameters was first proposed in [70].

There are different versions of the gradient descent algorithm based on how frequently the loss is computed. In the first version, the loss is computed after forward propagating on the complete dataset, followed by the backward propagation. This is commonly referred to as batch gradient descent, and although this procedure gives the true gradient and hence

better parameters $\mathbf{w}$, it is computationally expensive. On the other hand, it has been shown that computing loss with a small number of samples from the dataset each time, results in parameters $\mathbf{w}$ that are comparable to the batch gradient descent parameters. The small number of samples chosen each time is referred to as a mini-batch. Once the network has sampled through the entire dataset, referred as an epoch, the dataset is shuffled and sampled again in mini-batches to update parameters. This shuffling allows the network to see a group of samples in different combinations and allows the network to learn the complete distribution of the dataset. Based on the mini-batch size, the gradient descent is referred as stochastic gradient descent (SGD) if mini-batch size is one; or mini-batch gradient descent if it is greater than one and smaller than the complete dataset size. Generally, the mini-batch is of the size of 16 to 512 samples.

To further accelerate the gradient descent optimization, a momentum update is performed as $\mathbf{w} \leftarrow \mathbf{w} - \alpha m$. In comparison to gradient descent update, the momentum update uses the weighted average of gradients across multiple iterations $m = \beta_1 m + (1 - \beta_1)\frac{\partial L_\mathbf{w}}{\partial \mathbf{w}}$, where β_1 is typically 0.9. Momentum can also be seen as the first moment of the gradient. By using the momentum, the learning rate α that was fixed in the gradient descent algorithm is now adaptive due to weighted average of gradients operation. More recently, Adam [71] algorithm proposed to employ the second moment of gradient in addition to the first and showed them to be more robust empirically. The Adam algorithm update is defined by $\mathbf{w} \leftarrow \mathbf{w} - \alpha\frac{m}{\sqrt{v}+\epsilon}$, where $v = \beta_2 v + (1 - \beta_2)\frac{\partial L_\mathbf{w}}{\partial \mathbf{w}}^2$, m is the momentum and ϵ is used for avoiding division by zero. The recommended values for the constants in the paper [71] are $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The Adam optimizer is used in all experiments of this thesis.

Training and Hyper-parameter Tuning

Generally, the task-related metrics such as F-score or DOA error discussed in Section 2.4.2 are different from the objective functions used during optimization such as MSE or CE loss. This is mainly because the objective functions have to be differentiable to use an optimization algorithm such as Adam. Custom task related metrics may not always be differentiable and hence not directly usable for optimization. Further, during the optimization of neural network parameters using an optimization algorithm, care has to be taken to avoid over-fitting to the dataset. Such over-fitting leads to lack of generalization and hence poor performance on an unseen dataset.

To overcome this over-fitting and produce generic network parameters, the dataset is generally divided into training and validation splits. Most often the validation split is about 10-20% of the training split size. During the optimization, which is also referred to commonly as network training or simply training, the optimization algorithm computes the best network parameters using only the training data. After each epoch, the task-related metrics are used to calculate the scores on the unseen validation split. The training is stopped if the task related metrics do not improve for a few epochs, and the network parameters that gave the best result on the validation split is used as the final parameters. This process is referred to as early stopping and makes sure that the network is not over-fitting on the training split.

To decide parameters such as the number of layers and neurons in each layer, a hyper-parameter tuning is carried out. Different combinations of the number of layers and neurons are used to train a network. The training is stopped using early stopping on

validation split. The combination with the best task related metrics is chosen as the optimal network parameters for a given dataset.

Regularization

Regularization aims at training networks that generalize well to unseen data. Apart from early stopping that avoids over-fitting of the network to training data, methods like dropout [72] and batch normalization [73] can also be employed as regularizers.

Dropout [72] is a simple but effective regularization algorithm, where each neuron is kept active with some probability p during training. The probability p is a hyper-parameter and can be tuned similarly to other network parameters. By switching off individual neurons randomly, the connected neurons relying on these switched off neurons learn similar information from other active neurons. This reduces the strong co-adaptation between any two neurons. This kind of dropout training can be compared to training multiple neural networks, and averaging their outputs to obtain the final results.

Batch normalization [73] is an additional step performed after each layer of the network. This step normalizes the output activations of the previous layer to zero mean and unit variance. This normalization of activations, help in faster convergence of optimization algorithm.

3.5 Recurrent Neural Network

The FC network processes features of one time-frame t given by $\mathbf{x}_t$ and produces the corresponding output. Most often in sequence data, such as text, audio or video, the information is spread across multiple time-frames. For example, when reading this thesis, each word is understood based on the understanding of the previous words. One workaround to use FC with sequence data is to concatenate multiple time-frames into a single feature vector before processing it. But this means the number of frames to concatenate has to be chosen, producing another parameter to tune in the network. Further, within the dataset, different classes have different temporal distribution. For example in SED, impact sound events like door closing are short in duration, whereas bird calls or car horns are much longer. It is difficult to learn such class-wise information with the feature concatenation method in FC networks.

A recurrent neural network (RNN) overcomes this shortcoming of FC architectures by keeping in memory the information of previous inputs. Assume an input feature sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_T]$, and target labels $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t, ..., \mathbf{y}_T]$ of the same length. A RNN operation at frame t is given by

$$\mathbf{h}_t = f(\mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{W}_x \mathbf{x}_t), \tag{3.2}$$

$$\mathbf{y}_t = \mathbf{W}_y \mathbf{h}_t, \tag{3.3}$$

where $\mathbf{h}_t$ are the hidden state parameters learned from the current input frame $\mathbf{x}_t$, and $\mathbf{h}_{t-1}$ is the previous hidden state learned from all the previous frames $[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{t-1}]$. $\mathbf{W}_h$, $\mathbf{W}_x$ and $\mathbf{W}_y$ are the individual weights for hidden state parameters, input features and outputs, respectively. Most often the activation function $f()$ is a tanh. From the two equations, we see that the current output $\mathbf{y}_t$ is influenced by both the current input $\mathbf{x}_t$ and previous output states $\mathbf{h}_{t-1}$.

In some sequences, the current output $\mathbf{y}_t$ can be better estimated with the knowledge of not only the previous inputs but also the future inputs. For example, in speech, the

articulation of a phoneme depends on both the previous and future phonemes. With this information, an ASR system can perform better with both previous and future phonemes, than using just the previous phonemes. RNNs can be adapted to support learning from both forward and backward directions, this class of RNNs are referred to as bidirectional RNNs [74] and are employed for SED [I-III], localization [IV] and SELDT [V,VI] tasks.

Although RNNs theoretically can model long-term temporal dependencies, due to the number of time-steps involved the gradients vanish by the time they propagate to the initial time-steps during optimization. This is commonly referred to as the vanishing gradients problem [75]. The popular solutions to this are the gated recurrent units (GRU) [76] and the long-short term memory (LSTM) [77] units. Each of these units consist of multiple gates tracking the temporal information and the final output is a result of these gate activations. During training, each individual gate learns to identify time-steps of the sequence to ignore, the time-steps with relevant information to store, and the time-steps to trigger an output. Currently, both these methods are widely used for sequence tasks such as ASR [78], audio and image captioning [79, 80], machine translation [81], SED [I-III][62], localization [IV] and SELDT [V,VI]. Specifically for the SED tasks, it was shown that the LSTM network in [62] outperformed the FC network in [69].

3.6 Convolutional Neural Network

The neurons in an FC network are each connected to every other neuron in the previous layer. These connections might be redundant for datasets such as images, where similar spatial structures repeat multiple times within a single image. For example, a spatial structure such as vertical or horizontal edge can occur more than once in an image. Convolutional layers overcome this redundancy in FC and learn the local spatial structures. Given a three-dimensional input feature $\mathcal{X}$ and weight kernel $\mathcal{W}$, the output $\mathcal{H}$ is obtained as

$$\mathcal{H}_{i,j,k} = f(b + \sum_{l} \sum_{m} \sum_{n} \mathcal{X}_{i-l,j-m,k-n} \mathcal{W}_{l,m,n}) \tag{3.4}$$

where b is the bias for kernel $\mathcal{W}$, and the activation function $f()$ is commonly ReLU. The depth k is equal to one for monochrome images, whereas for color images of RGB (red, green and blue channels) format it is equal to three. Similarly, when treating a single-channel magnitude spectrogram of audio as the input image for the convolutional layer, k is equal to one, whereas, for multichannel magnitude spectrograms k is equal to the number of channels. Generally, the kernel size is much smaller than the input image size. As seen from the equation, the same kernel weights $\mathcal{W}$ are shared across the image thus reducing the redundancy of connections. Hence convolutional layers are memory-efficient, faster, and have a lower number of weights than the FC networks. Each kernel $\mathcal{W}$ is trained to learn one particular local structure and will detect if this structure repeats within an image or in another image. For example in a single-channel magnitude spectrogram input containing two identical bird-calls that are temporally separated and shifted in frequency, both the bird-calls can be detected by a single kernel that has learned the bird-call structure. This property of the convolutional layer to detect similar local structure across the image is called shift-invariance. Further, each kernel of the convolutional layer produces an output of $\mathcal{H}$ that is often referred to as the feature map. For complex classification tasks, these convolutional layers are made up of multiple kernels, and a corresponding number of feature maps. The kernels are also called filters in some literature.

A convolutional neural network (CNN) comprises of multiple convolutional layers each followed by an optional max pooling operation. The max pooling operation is performed to reduce the dimensionality of the input feature. Max pooling is performed by first dividing the output $\mathcal{H}$ into multiple two-dimensional regions of fixed-dimensions tuned by the user, and choosing the maximum value in each of the regions. This is similar to the downsampling operation in two-dimensions. The max pooling can also be considered as a generalizing step that enables learning relevant information from a high-resolution input. In the case of SELDT, where the frame-level activity of sound events is required, the max pooling operation is performed only in the frequency axis keeping the number of frames in the output unchanged from the input. Finally, as the classification layer, often the feature map of the last convolutional layer is fed as input to an FC layer which produces the class-wise probabilities. Such CNN networks have been used extensively for classification tasks that do not need temporal information, such as sound event classification [82].

3.7 Convolutional Recurrent Neural Network

The FC, convolutional and recurrent layers discussed so far each learn complimentary information from one another. The FC layers are good at frame-wise classification but are poor in modeling temporal structures and the number of parameters increases with the size of the input. Convolutional layers are good at learning shift-invariant features from two- or three-dimensional input. Further, the shared kernels mean the number of parameters is significantly smaller than an FC for the same input size. On the other hand, they fail at modeling temporal structures similar to FC. The RNNs are good at learning temporal information from a two-dimensional input but have a similar problem as FC with respect to the number of parameters. Thus combining these complimentary blocks can result in a powerful classifier for sequence data.

A CNN can be used to learn shift-invariant features from a high-resolution input such as magnitude spectrogram. This results in using fewer parameters than an FC or RNN and also produces robust features that contain the abstract information in a much lower dimension as a result of multiple max pooling operations. The three-dimensional feature map output of the last convolutional layer is then reshaped into two-dimensions of temporal and feature axes and fed to an RNN to learn the temporal information. The RNN is mostly made up of either GRU or LSTM units and can be bidirectional in nature. The two-dimensional output of RNN is comprised of the temporal and feature axes and is fed to an FC layer to produce frame-wise activity, for example of the sound events. The weights of this FC layer are shared across each time-frame output of RNN. Employing the RNN and FC on the low-dimension CNN feature map helps to keep the total number of network parameters small. This joint architecture is referred to as convolutional recurrent neural network (CRNN) and has been used successfully on many audio tasks such as ASR [83], music genre classification [84], music emotion recognition [85], sound event classification [4], SED [II,III], localization [IV] and SELDT [V,VI]. Specifically, for the SED task the CRNN, with GRU units for RNN in [II,III] outperformed the RNN only network with LSTM units in [I][62].

4 Sound Event Detection

In this chapter, we first present a brief discussion on the related works for the SED task. Thereafter, we present our work on multichannel SED proposed in [I-III]. Here, for each publication, we provide the motivation, discuss the method, present the evaluation results and finally provide the summary of our contributions and limitations of the proposed method.

4.1 Related Works

In the literature, the SED task has typically been approached as a supervised learning task that maps frame-wise input features to frame-wise sound event activity. Some of the early SED methods [64] recognized only the dominant sound event among the overlapping sound events and its corresponding onset-offset times using a GMM-HMM classifier. However, this approach is not suitable for applications that require the detection of multiple sound events occurring simultaneously. During the time of writing [I] most methods were proposing to perform polyphonic SED, i.e., the task of recognizing multiple overlapping sound events and their respective temporal activities.

Initially, few methods proposed to exploit the prior knowledge of the acoustic scene to build models that are scene-specific. This was developed with a similar idea as the language model in ASR, i.e., not all phonemes occur in all languages. Similarly, not all sound events occur in all acoustic scenes, and hence we can exploit this information to produce better SED. This class of methods was referred to as scene-dependent SED [86]. However, such scene information is not always available, hence more recent methods have been studying scene-independent SED.

Some of the common classifiers employed for the polyphonic SED task are GMM-HMM [86], non-negative matrix factorization (NMF) [87], FC [69], CNN [88, 89], and RNN [62, 90] networks. In [86] the overlapping sound events were recognized using multiple restricted Viterbi passes with the GMM-HMM model. To recognize the overlapping sound events better, [87] proposed to use an NMF algorithm as a preprocessing step to obtain multiple streams of source separated audio. A GMM-HMM model similar to [64] was then used on each of the separated audio streams to obtain the temporal activity of overlapping sound events.

With the advent of deep learning techniques, several DNN based methods were proposed for the polyphonic SED task. Cakir et. al. [69] performed SED as a multiclass multilabel classification task with an FC network. To provide contextual information, each input frame to the FC network was obtained by concatenating multiple time-frames of the feature. This method performed better than the previous best SED method [87]. Among generative classifiers, such as GMM, individual GMMs are trained for each sound class.

During inference, the sound class is assigned based on the GMM with the best likelihood results. In a similar way, multiple FC classifiers were trained in [91] for each sound class in the dataset. During inference, the combined results of the multiple single-class FC classifiers were used for SED. The SED performance of this was compared with a multiclass multilabel FC classifier. The results showed that the multiclass multilabel approach performed marginally better than a multiple single-class approach.

The RNNs are a class of DNN that were developed to model sequence data, such as text, speech, and audio, and hence are best suited for the SED task. Multiple LSTM layers were employed in [62] to perform SED as a multiclass multilabel classification and produced better results than the FC method in [69]. At the time of writing [I], multiple LSTM layers network [62, 90] was the state-of-the-art for the SED task.

All these SED methods were using some form of a spectral feature such as mel-frequency cepstral coefficients and log mel-band energies. Further, all these methods were using single-channel audio input. Recognizing multiple overlapping sound events with single-channel audio can be a challenging task. These overlapping sound events can potentially be recognized better with multichannel audio. The first SED method using multichannel audio was proposed in [67]. This method combined the classification likelihood scores across channels to perform SED. Although it used multichannel audio, it did not exploit the spatial information from it.

In this regard, the first method employing spatial features extracted from multichannel audio for polyphonic SED was proposed in [I] and described in Section 4.2. Further studies on different acoustic features and DNN architectures for multichannel SED were presented in [II] and [III], described in Section 4.3 and Section 4.4 respectively. Finally, the SED performance with respect to the number of audio channels; and the relation between the DNN size and the number of overlapping sound events in an acoustic scene is studied in [III] and described in Section 4.4.

4.2 Sound Event Detection in Multichannel Audio Using Spatial And Harmonic Features

The human auditory system has been exploiting the binaural audio cues to recognize confidently multiple overlapping sound events. Motivated by this, we proposed to employ binaural audio instead of the traditionally used single-channel audio in [I] to improve the recognition of overlapping sound events. Additionally, inspired by how the human auditory system processes binaural audio for recognizing overlapping sound events, we proposed spatial and harmonic features extracted from binaural audio that provide similar information for the polyphonic SED task.

4.2.1 Method

The proposed method is composed of two stages: the sound representation and a multiclass multilabel classifier implemented using RNN layers with LSTM units. Three different acoustic features motivated by the human auditory system were proposed in the sound representation stage. All three features were extracted using windows with identical hop-length. Hence the number of time-frames for an audio recording of a given length is constant across the features used. The studied features and its variations are presented in Table 4.1. As the first feature, the log mel-band energy (*mbe*) (see Section 3.1 for details) extracted from each channel of the binaural audio was used. Extracting *mbe* from the

binaural audio was motivated by the interaural intensity difference (IID) used by the human auditory system for spatial localization of sound sources [92]. The DNNs which are capable of performing linear operations, including the difference, can obtain this IID information from the binaural channel *mbe*. A 40-band *mbe* feature is extracted from 40 ms windows with 50 % overlap using Hamming window from each channel of the binaural audio.

Human listeners have been identifying different sound events using perceptual features such as pitch [93]. Studies in [94] showed that using the pitch in addition to MFCCs improved the non-speech environmental sound detection performance. Motivated by this, in [I] we proposed to use the dominant frequencies and their respective magnitudes (referred together as *dom-freq* hereafter) estimated from both the channels as our second feature. Since sound events are not always harmonic, pitch values for non-harmonic events do not exist. Hence we simply choose *dom-freq* values from the librosa implementation [95] of thresholded parabolically-interpolated STFT [96] in the 100-4000 Hz range from each of the binaural channels using 40 ms windows and 50 % overlap. Since the dataset studied, TUT-SED 2016 (Section 2.4.1), had a maximum of three temporally overlapping sound events we choose the top three dominant frequency values and their respective magnitudes for each time-frame (referred together as *dom-freq*3 hereafter) as the second set of perceptual features.

Sound source localization is an important cue used by the human auditory system to recognize multiple overlapping sound events better. A strong cue for localization at low frequencies is the interaural time delay (ITD) [92]. Motivated by this, we implemented the ITD with the binaural audio using time difference of arrival (TDOA) extracted in different frequency bands. Sound events typically have different spectral range, with some events occurring at lower frequencies, others at high frequencies, and some across the entire spectrum.. By estimating the TDOA in different frequency bands, an isolated sound event with wide band spectrum will have the same TDOA value in all the frequency bands. On the other hand, two temporally overlapping sound events which are distributed locally in different frequency bands will have different TDOA values. In this study, TDOA was calculated in five mel-bands. In each of the mel-band b, the generalized cross-correlation with phase-based weighting (GCC-PHAT) [97] was calculated as

$$R_b(\Delta_{12}, t) = \sum_{k=0}^{N-1} H_b(k) \frac{X_1(k,t) \cdot X_2^*(k,t)}{|X_1(k,t)||X_2(k,t)|} e^{i2\pi k \Delta_{12}/N}, \tag{4.1}$$

where $X(k,t)$ is the DFT coefficient at time-frame t and frequency bin k. The subscripts $_1$ and $_2$ represent the binaural channel numbers. N is the total number of frequency bins in the DFT. $H_b(k)$ is the magnitude response of the b-th mel band and $\Delta_{12} \in [-\tau_{\max}, \tau_{\max}]$, where $\tau_{\max} = 30$ is the maximum sample delay for a sound wave to travel between binaural microphones. Finally, the TDOA values for each band b and time-frame t is obtained by picking the respective peak from the GCC-PHAT using,

$$\tau(b,t) = \underset{\Delta_{12}}{\operatorname{argmax}} \{R_b(\Delta_{12}, t)\}. \tag{4.2}$$

To accommodate sound events of different lengths, the TDOA values are calculated using three different windows of length 120 ms, 240 ms, and 480 ms. Hence, picking five TDOA values in the respective five mel-bands for each of the three windows results in 15 TDOA values per time-frame (referred to as *tdoa*3 hereafter).

Table 4.1: Binaural audio acoustic features proposed for sound event detection.

Feature Name	Length	Description
mbe	40	Log mel-band energy extracted on a single-channel of audio
dom-freq	2	Most dominant frequency value and periodicity extracted on a single-channel
*dom-freq*3	6	Top three dominant frequencies and periodicity values extracted on a single-channel
tdoa	5	Median of multi-window TDOA's extracted from stereo audio
*tdoa*3	15	Concatenated multi-window TDOA's extracted from stereo audio

TDOA values from real-life audio are noisy in general, even more so when estimated in small window lengths. To overcome this, the median TDOA value across the three window lengths was used as the second set of TDOA features (referred as *tdoa* hereafter).

As the second stage of the method, a DNN comprising multiple LSTM layers was employed to map the acoustic features to SED using multiclass multilabel classification. Specifically, we used two hidden layers of LSTM with 32 units each, and the output layer had a number of units equal to the number of classes in the dataset with sigmoid activation layer to enable multiclass multilabel classification. During training, the respective feature-combinations were concatenated along feature-axis and non-overlapping sequences of length 25 frames were provided as input to the network. A one-hot encoded reference annotation from the dataset was used to compute the cross-entropy loss with the predicted output of the classifier. Adam optimizer was used along with early stopping to reduce overfitting. During inference, the class-wise probability is thresholded with a constant 0.5 value. A sound event is marked active if it is greater than 0.5 and otherwise inactive.

4.2.2 Evaluation

The method [I] is evaluated on the TUT-SED 2016 development dataset, which contains binaural recordings for two contexts – home and residential area, with 11 and 7 sound event classes respectively (see Section 2.4.1 for more details). The performance is evaluated with the SED metrics F-score and ER calculated for non-overlapping one-second segments and presented in Table 4.2. The baseline method [98] uses mel frequency cepstral coefficients (MFCC), its first and second moments to train a GMM classifier. A separate positive GMM model is trained for every sound event class, and a corresponding negative GMM model is trained for each class using the rest of the audio. During testing, the likelihood ratio between the positive and negative GMM model for each individual class is used for activity detection. This baseline method achieves an average ER of 0.91 and F-score of 23.7 %. In comparison, the proposed LSTM method with just the single-channel *mbe* feature achieves a better F-score of 32.9 % for the same ER as the baseline. This single-channel *mbe* feature was created by averaging the binaural channels of the audio. In comparison to single-channel features, all the proposed binaural feature combinations apart from $mbe; tdoa3$ and $mbe; tdoa3; dom\text{-}freq3$ gave better F-score. This suggests that binaural features are helpful for the polyphonic SED task.

4.2.3 Contributions and Limitations

The proposed method in [I] is the first method to exploit spatial and harmonic features motivated by the human auditory system to perform SED using binaural audio. The results suggest that using binaural audio improves polyphonic SED performance. In fact, the proposed LSTM method with the binaural *mbe* and $mbe; tdoa; dom\text{-}freq$ features won the first and second position respectively in the real-life SED task of the IEEE AASP

Table 4.2: The evaluation scores of proposed CRNN for different feature combinations on TUT-SED 2016 development dataset. The different features employed are presented in Table 4.1. The combination *mbe*; *tdoa*; *dom-freq* represents the log mel-band energies from the two channels, the most dominant frequency and respective magnitude values extracted from the two channels, and the time difference of arrival calculated between the two channels. The best scores for the respective metric are highlighted.

	Home		Residential area		Average	
	ER	F (%)	ER	F (%)	ER	F (%)
Single-channel features						
GMM trained on *mfcc*; Δ; ΔΔ [98][99]	0.96	15.9	0.86	31.5	**0.91**	23.7
mbe	**0.94**	**27.4**	0.88	38.3	**0.91**	32.9
Binaural feature combinations						
mbe	1.03	25.4	0.84	45.9	0.93	**35.6**
mbe; *dom-freq*	1.03	24.9	0.93	40.9	0.98	32.9
mbe; *dom-freq*3	0.97	26.6	0.88	41.7	0.92	34.2
mbe; *tdoa*	1.01	24.4	**0.82**	**46.4**	**0.91**	35.4
mbe; *tdoa*3	0.96	24.9	0.86	38.5	**0.91**	31.7
mbe; *tdoa*3; *dom-freq*	0.97	25.7	0.85	43.1	**0.91**	34.4
mbe; *tdoa*3; *dom-freq*3	0.99	26.5	0.91	35.2	0.95	30.9
mbe; *tdoa*; *dom-freq*	0.98	24.7	0.87	43.8	0.92	34.2
mbe; *tdoa*; *dom-freq*3	**0.94**	26.3	0.89	40.5	**0.91**	33.4

challenge DCASE 2016 [99]. Further, among the 17 competing methods in the challenge, the combination of binaural *mbe* and LSTM classifier was the only method that performed better than the benchmark[1]. Making the proposed method the state-of-the-art.

Although the results suggest that using binaural audio features instead of single-channel improves SED performance, it is difficult to conclusively claim the binaural features employed in this study are superior to single-channel features since the evaluated TUT-SED 2016 dataset size is small (about an hour long). Further, the spatial features studied are highly hand-crafted requiring many variables to be tuned. Such features cannot be used as generic features and will have to be re-tuned based on the sound event classes and their distribution in the respective dataset. Additionally, the TDOA values in the higher frequency bands might be ambiguous, hence the frequency bands have to be carefully chosen. Finally, the current representation of concatenating different feature classes along feature-axis does not allow for optimum learning of multichannel information. A better feature representation that enables a classifier to learn the intra- and inter-channel features in a more optimal way are addressed in the next section.

4.3 Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network

Motivated by the state-of-the-art SED performance obtained using the binaural features and multilayered LSTM network in [I], we continued to explore multichannel DNN methods and feature generalization in [II]. Specifically, we proposed to use low-level

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio#results

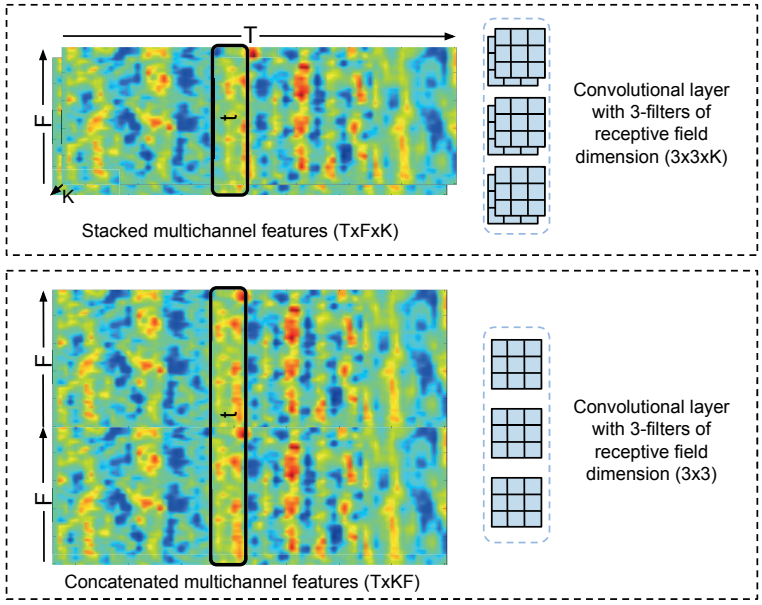


Figure 4.1: Stacked vs. concatenated multichannel feature representation as input for convolutional layers. Where K denotes the number of channels, T and F denotes the number of time-frames and feature-length respectively.

features in place of highly handcrafted features to avoid tuning features individually for different datasets. For example, the TDOA features used in [I] are highly handcrafted with tunable variables such as the number of bands, frequency ranges of the respective bands, the number of peaks per band, and the number of different window resolutions. Instead, if we can use a lower level feature with fewer tunable parameters such as the GCC-PHAT (referred as *gcc* hereafter) from which the TDOA feature is extracted, and show that a network learns similar information from both TDOA and *gcc*, then this will avoid handcrafted feature extraction and make the features dataset independent. Analogously, we propose replacing the handcrafter dom-freq features with a low-level autocorrelation feature $dom\text{-}freq$ with low-level autocorrelation feature (referred as *acr* feature hereafter) is also proposed.

One drawback of the recurrent layers model used in [I] is that the temporal modeling is done directly on the input acoustic feature (*mbe*, $dom\text{-}freq$, and *tdoa*) sequence. However, higher-level modeling of the information in the input acoustic feature can disentangle the information in the feature and enable the recurrent layers to learn better [100]. This was first shown in [83] using CRNN, where multiple convolutional layers were first used to disentangle the input feature sequence, followed by temporal modeling of the convolutional layer output using RNNs, and finally mapped to the output classes using a fully connected layer. At the time of writing [II], CRNN methods were state-of-the-art for multiple audio domain tasks such as automatic speech recognition [83], and SED [61]. Motivated by the CRNN performance we adapted it to support multichannel audio feature input, and multiple feature classes in [II]. Finally, to conclusively claim the benefits of binaural features for polyphonic SED, the proposed method in [II] was evaluated with the 19-hour long TUT-SED 2009 dataset.

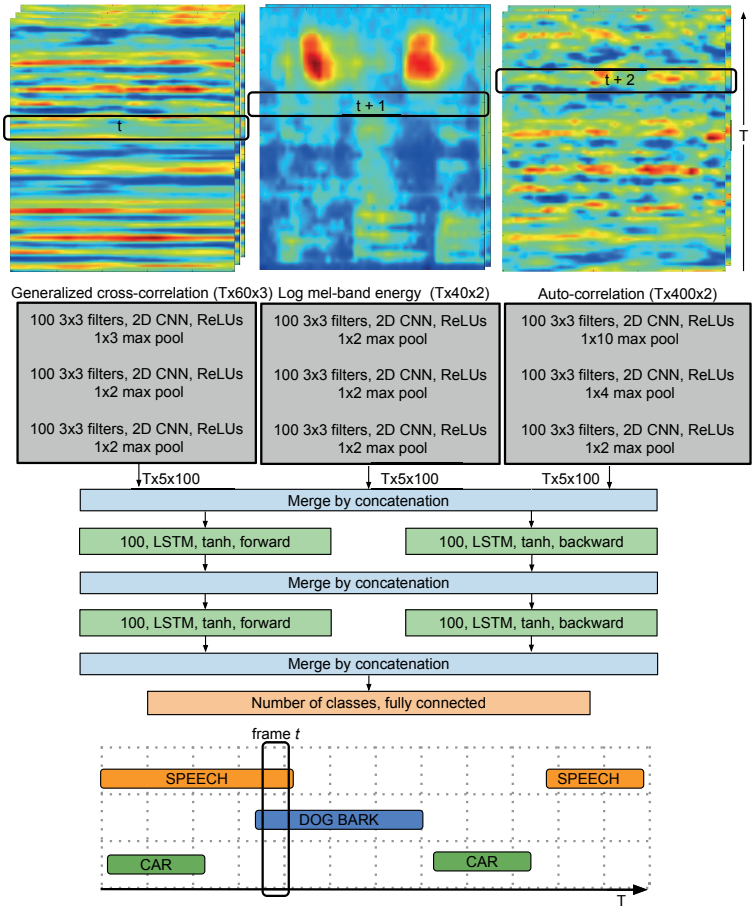


Figure 4.2: Proposed architecture of convolutional recurrent neural network (CRNN) for multichannel audio features and multiple feature classes. ©2017 IEEE.

4.3.1 Method

The proposed method in [II] is composed of two stages: the sound representation and multiclass multilabel classification using CRNN. Similar to [I] we continue to use the *mbe*, *dom-freq*3 and *tdoa*3 features. Additionally we propose two low level features: *gcc* replacing *dom-freq*3 and *acr* replacing *tdoa*3. *gcc* is calculated using Equation 4.1 in a single band. A total of 60 *gcc* values are picked in lags $\Delta_{12} \in [-29, 30]$ in three windows of length 120 ms, 240 ms, and 480 ms to accommodate sound events of variable length, amounting to 180 *gcc* values per time-frame. Similarly, we replace the *dom-freq*3 with *acr* values from which the perceptual feature pitch used by the human auditory system is inferred. *acr* is calculated using time-domain auto-correlation using 40 ms windows and choosing 400 correlation values in the range of 107.5-4410 Hz. Both the number of *acr* and *gcc* values per time-frame were chosen such that they are easily factorizable during max-pooling operation in CNNs.

As the input feature representation, we could use the multiple feature classes concatenation along feature-axis similar to [I]. But to enable the first convolutional layer of the CRNN to learn better from the multichannel and multiple feature classes, we split the features

Table 4.3: The evaluation scores of proposed multichannel CRNN and baseline single-channel CRNN for TUT-SED 2009 and 2016 datasets. The best scores for the respective metric are highlighted. ©2017 IEEE.

Feature combination	TUT-SED 2009		TUT-SED 2016	
	ER	F	ER	F
Single-channel				
CRNN baseline [61]	0.49	68.8	**0.93**	31.3
mbe	0.49	68.0	1.03	29.7
Binaural				
mbe-concat	0.44	70.3		
mbe	**0.43**	71.1	0.99	32.3
mbe + *tdoa*3	0.45	70.9	0.95	**35.8**
mbe + *gcc*	0.44	71.1	0.95	34.6
mbe + *dom-freq*3	**0.43**	**71.7**	0.98	32.8
mbe + *acr*	0.44	71.2	0.98	33.8
mbe + *tdoa*3 + *dom-freq*3	0.44	71.0	1.01	33.3
mbe + *gcc* + *acr*	0.45	70.9	0.99	33.6

as follows. A T frames 40-band *mbe* feature from the two binaural channels is stacked to form a volume of dimension $T \times 40 \times 2$ as shown in Figure 4.1. The 2D convolutional layers by design are built to learn from these volumes, i.e., the number of channels in each of the convolutional layer filter is equal to the number of input feature-channels. During training each individual filter-channel learns information from the corresponding feature-channel, and the weighted combination of these channel-wise filters produces the output. This enabes the convolutional layer to learn both channel-specific and multichannel information. We compare the performance of using such stacking with respect to concatenating multichannel features as shown in Figure 4.1 during evaluation. Similar to *mbe* the five band *tdoa*3 peaks in three multiresolution windows are stacked to $T \times 5 \times 3$ and the low-level feature *gcc* is stacked as $T \times 60 \times 3$. Finally, the three dominant frequencies and their respective magnitudes of *dom-freq*3 feature are treated as separate layers of dimension $T \times 3 \times 8$ and the 400 values of *acr* are stacked as $T \times 400 \times 2$ dimension volume.

The overall structure of the tuned CRNN for the three features - *gcc*, *acr* and *mbe* are shown in Figure 4.2. A separate set of convolutional layers are used to learn shift-invariant features from each of the three feature volumes. This enables the convolutional layers to learn feature-specific filters. When using handcrafted features *tdoa*3 and *dom-freq*3 only one convolutional layer of 100 filters is used without max-pooling operation. In both the cases, the max-pooling operation is only performed on the feature axis thus keeping the time resolution of the output equal to input (T). The output activation from the three convolutional layer blocks are concatenated along feature-axis and fed to a multilayered recurrent network to learn the temporal context of sound events, and finally, this is mapped to the multilabeled output classes using a fully connected network with sigmoid activation. The CRNN architecture is jointly trained with the Adam optimizer using cross-entropy loss for 500 epochs. Early stopping was used to stop overfitting if the F-score of SED did not change for 50 epochs. Batch normalization was used after every convolutional layer, and a fixed 50 % dropout was used in the convolutional and recurrent layers.

Table 4.4: The F-scores of the proposed CRNN for different contexts of TUT-SED 2009 dataset. The feature combination obtaining the best F-score for the respective context is highlighted. ©2017 IEEE.

Feature combination	Indoor						Outdoor		
	Basketball	Bus	Hallway	Office	Car	Restaurant	Shop	Beach	Street	Track and Field
Single-channel										
mbe	79.7	52.6	59.1	81.8	78.2	80.7	62.4	56.5	60.3	70.1
Binaural										
mbe	82.2	56.5	66.6	83.3	81.5	83.1	63.3	59.5	66.0	70.2
mbe + *tdoa3*	82.8	58.7	66.0	80.8	79.2	81.2	64.7	**60.9**	**66.9**	68.8
mbe + *gcc*	81.9	58.9	65.3	80.0	81.2	81.3	65.3	60.4	66.3	**72.6**
mbe + *dom-freq3*	**83.7**	**60.5**	**67.8**	**84.6**	80.8	81.8	64.6	60.7	66.6	67.6
mbe + *acr*	82.9	58.6	63.8	83.6	**83.4**	**82.3**	**65.5**	60.4	65.8	69.0
mbe + *tdoa3* + *dom-freq3*	83.0	59.4	67.5	83.9	78.6	79.9	65.1	60.5	65.0	70.0
mbe + *gcc* + *acr*	82.8	59.1	66.8	82.2	79.4	80.4	64.8	60.4	66.1	68.5

4.3.2 Evaluation

The proposed method was evaluated on the publicly available TUT-SED 2016 dataset (see Section 2.4.1) of about one-hour duration, and TUT-SED 2009 dataset of about 19 hours duration (see Section 2.4.1). The SED metric results obtained are presented in Table 4.3. As the baseline, we used the existing state-of-the-art CRNN method proposed in [61] for single-channel SED. The baseline method uses identical single-channel *mbe* features and maps them to the SED activity using a CRNN architecture with hyper-parameters different from the proposed method. In comparison, the proposed architecture with single-channel *mbe* features obtains similar performance on TUT-SED 2009, but the performance drops with TUT-SED 2016. We believe that the drop in performance for TUT-SED 2016 is due to dataset size. Since it is relatively small, the variation in the SED scores obtained were relatively high.

With regard to input representation for binaural features, from Table 4.3 we see that the proposed CRNN performs better with stacked *mbe* feature than the concatenated (*mbe-concat*). Compared to the single-channel *mbe* feature using the binaural features with the proposed CRNN was observed to improve the ER and F-score across datasets. Using binaural stacked features instead of single-channel features gave an absolute F-score improvement of at least 2.9 % for TUT-SED 2009 (*mbe* + *gcc* + *acr*) and 2.6 % on TUT-SED 2016 dataset (*mbe*). With these results, we can confidently claim that binaural (multichannel) information helps polyphonic SED, and the proposed network is truly learning the relevant binaural information from the input features.

From Table 4.3 we see that replacing *tdoa*3 with *gcc* or *dom-freq*3 with *acr* yields comparable SED performance across datasets. This is a significant result, showing that the network can learn the information equivalent to handcrafted features from the low-level features directly, and thereby making the feature extraction process dataset independent.

Table 4.4 presents the SED metrics obtained for different contexts of TUT-SED 2009 dataset. In general, it is observed that perceptual features *acr*/*dom-freq*3 helps in indoor scenarios such as basketball court, bus, and hallway. Whereas spatial features *gcc*/*tdoa*3 helps in outdoor scenarios. We hypothesize that the indoor scenarios are small-spaced and reverberant, hence the information in the spatial features might be ambiguous. This resulted in the network to learn more from the perceptual features that are affected less with reverberation. On the other hand, for the outdoor scenarios that are open areas with little reverberation, the sound events are located significantly further from each other resulting in the network learning more from the spatial features. This also explains why in the Table 4.3 the *mbe* and *dom-freq*3 feature combination gave the best results for TUT-SED 2009 whereas the *mbe* and *tdoa*3 gave the best results for TUT-SED 2016 dataset. The TUT-SED 2009 has more indoor context recordings, while TUT-SED 2016 has an equal number of indoor and outdoor context recordings.

4.3.3 Contributions and Limitations

In [II], we proposed a CRNN architecture that supports multiple feature classes. We showed that stacking the multichannel features enables the first convolutional layer of CRNN to learn the multichannel information better than simple feature concatenation. To the best of our knowledge, this is the first work which uses stacked multichannel audio features as input to a deep learning method. The proposed CRNN was shown to learn information equivalent to handcrafted features from naïve low-level features, thereby making the feature extraction dataset independent. The performance of the

proposed method was evaluated on two datasets. The SED performance results obtained conclusively show that using binaural features in addition to single-channel features improves polyphonic SED.

The proposed CRNN and *mbe* feature won both the first (single-channel) and second (binaural) position in the DCASE 2017 real-life SED task[2] making it the state-of-the-art SED method. To support reproducibility of research, the winning architecture is made publicly available[3].

In [II], although we showed that the binaural audio helps polyphonic SED, we did not study the SED performance in different polyphonic scenes such as a quiet scene (little polyphony) vs sound intensive scene (higher polyphony). No studies were performed to understand the performance at time-frames with different number of overlapping sound events. Further the results from [II] raise questions about whether SED performance continues to improve with a higher number of channels. We answer some of these questions in the next section.

4.4 Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features

In [III], we study the SED performance in identical scenes using a single-channel, binaural and four-channel FOA format audio. These audio formats with different number of channels were all generated from the FOA recordings. The omnidirectional W channel of FOA is used as the single-channel recording, binauralized FOA recordings using real head-related transfer functions (HRTF) are utilized as the binaural recordings, and FOA as the multichannel recording. We limit the number of channels to four since most consumer devices, especially recording 360° audio, use four-channel FOA format audio.

To study the SED performance in acoustic scenes with different polyphony, we use the synthetic ANSYN dataset (see Section 2.4.1) since it is difficult to collect a controlled dataset with varying polyphony in real-life. For this paper, we used the $O1$ and $O3$ subsets from ANSYN and generated a new subset $O6$ that has a maximum of six temporally overlapping sound events.

As a classifier, we continued to use the CRNN proposed in [II]. Further, we proposed to use one layer of a 3D convolutional layer as the input layer of CRNN (referred to as C3RNN hereafter) to enable the respective filters to learn both inter- and intra-channel information from the audio features.

4.4.1 Method

Similar to the methods in [I, II], the proposed method in [III] is composed of two stages: sound representation and multiclass multilabel classifier. In the sound representation stage, we extract *mbe* and *gcc* from the input audio. For a sequence length of T frames the 40-band *mbe* feature is of a general $T \times 40 \times K$ dimension, where the number of channels $K = 1, 2, 4$ for single-channel (*mbe-mono*), binaural (*mbe-bin*) and multichannel FOA (*mbe-ambi*) formats respectively. Similarly, the 60 *gcc* values in three multiresolution

[2] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-sound-event-detection-in-real-life-audio#results

[3] https://github.com/sharathadavanne/sed-crnn

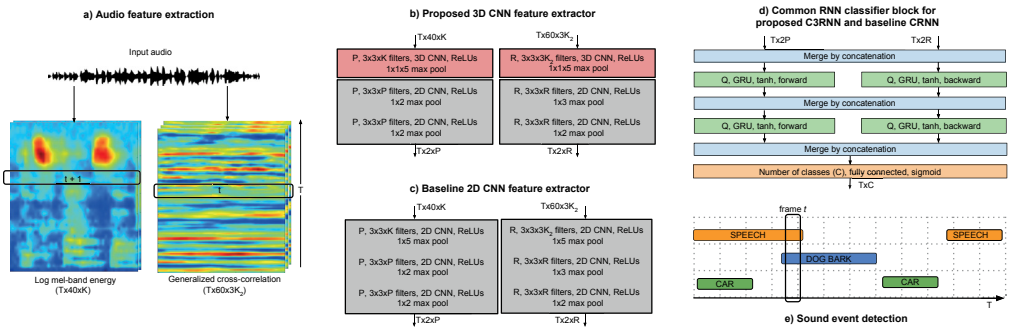


Figure 4.3: Convolutional recurrent neural network architectures for multichannel sound event detection. The proposed C3RNN (a+b+d+e) method employs a 3D CNN input layer, whereas the baseline CRNN (a+c+d+e) method employs only 2D CNN layers. ©2018 IEEE.

windows of length 120, 240, and 480 ms are of general dimension $T \times 60 \times 3\binom{K}{2}$, where 3 is the number of multiresolution windows and $\binom{K}{2}$ is the number of channel pairs for K-channels. Since *gcc* can only be extracted for channel pairs, *gcc* is defined only for binaural (*gcc-bin* of dimension $T \times 60 \times 3$) and for FOA format (*gcc-ambi* of dimension $T \times 60 \times 18$).

As the classifier, we use the state-of-the-art CRNN [II] and C3RNN which is identical to CRNN but with a 3D convolutional layer as the first input layer. Both the classifiers are visualized together in Figure 4.3. Although the filter dimensions of the first 2D convolutional layer in CRNN and first 3D convolutional layer in C3RNN are identical ($3 \times 3 \times K$ for *mbe* branch and $3 \times 3 \times \binom{K}{2}$ for the *gcc* branch), the key difference between them is in the convolution pattern. In CRNN each channel of the filter is convolved with only one corresponding input feature channel, whereas in C3RNN each channel of the filter is convolved with all channels of the input feature. In theory, the 3D convolutional layer filter is a hybrid of both stacking and concatenating the multichannel input feature, i.e., individual filters for each input feature channel additionally learn from corresponding feature channels.

The two networks are trained for 1000 epochs using Adam optimizer and cross-entropy loss between the reference and predicted sound class activities. Early stopping is used to stop overfitting of the network to training data. The training is stopped if the ER metric on the test split does not improve for 100 epochs. Batch normalization is used after each convolutional layer, and a fixed dropout is used for all convolutional and recurrent layers.

4.4.2 Evaluation

Independent of the feature used, for a given dataset (ANSYN $O1$, $O3$, $O6$ and TUT-SED 2017) the hyperparameters remained the same. A 128-frame sequence length, 32 batch size, and 0.35 dropout gave the best results across ANSYN subsets and features combinations. A sequence length of 256 frames, the batch size of 128 and 0.2 dropout gave the best results for TUT-SED 2017 dataset. We continued to use a similar number of layers as in [II] but tuned the number of units in each layer for each dataset. In case of TUT-SED 2017 the number of convolutional filters in each layer for *mbe* (P in Figure 4.3) and *gcc* (R in Figure 4.3) branch was equal to 64, and the number of GRU units Q was equal to 64. Similarly for ANSYN $O1$ $P = Q = 8$ and $R = 16$, for $O3$ $P = Q = 16$ and

Table 4.5: The SED evaluation scores of proposed C3RNN and baseline CRNN methods for ANSYN subsets. The best scores for the respective metric are highlighted. ©2018 IEEE.

	*O*1		*O*3		*O*6	
C3RNN	ER	F	ER	F	ER	F
mbe-gcc-ambi	0.11	92.2	0.18	82.5	0.17	84.1
mbe-gcc-bin	0.12	91.6	0.20	79.8	0.24	77.2
mbe-ambi	**0.09**	93.7	**0.16**	**83.8**	**0.16**	**85.4**
mbe-bin	0.10	**93.8**	0.18	81.8	0.22	78.5
mbe-mono	0.10	91.9	0.17	81.8	0.26	77.9
CRNN	ER	F	ER	F	ER	F
mbe-gcc-ambi	0.11	91.1	0.19	81.6	0.19	83.5
mbe-gcc-bin	0.12	92.3	0.21	78.8	0.26	79.0
mbe-ambi	**0.10**	92.8	**0.18**	**82.5**	**0.17**	**83.7**
mbe-bin	0.11	**93.6**	0.19	79.3	0.23	79.5
mbe-mono	0.12	91.9	**0.18**	80.6	0.28	78.3

$R = 32$, and for $O6$ $P = Q = 32$ and $R = 64$. This correlation of an increasing number of convolutional and recurrent units with a higher number of overlapping sound events in the tuned networks show that bigger networks are required to recognize sound intensive scenes.

The SED metric scores obtained for the ANSYN subsets with CRNN and C3RNN are presented in Table 4.5. Analyzing the C3RNN performance with *mbe* features, we observe that the SED performance across single, binaural and multichannel features are comparable for $O1$ subset. However, with the increase in polyphony ($O3$ and $O6$) the SED performance with binaural (*mbe-bin*) and FOA (*mbe-ambi*) format audio is better than single-channel (*mbe-mono*). A similar pattern with *mbe* features is also observed with the CRNN classifier. CRNN and C3RNN yielded comparable results. The main advantage of using the C3RNN over CRNN was observed in the training speed as shown in Figure 4.4. In terms of the number of epochs taken to train, the C3RNN achieved better ER with a lower number of epochs. Although this does not reflect on the number of computations, the C3RNN takes more computational power than CRNN because each

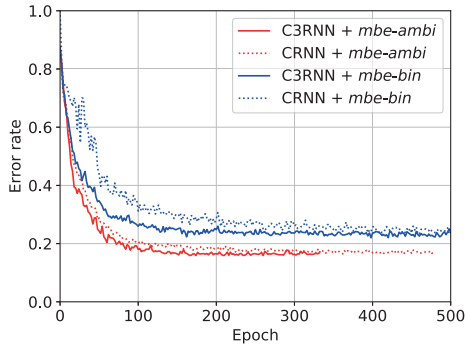


Figure 4.4: The learning curve for the proposed C3RNN and baseline CRNN methods using ambisonic and binaural *mbe* features extracted from ANSYN $O6$ subset. The proposed C3RNN achieves better error rate in fewer epochs for both ambisonic and binaural features. ©2018 IEEE.

Table 4.6: The frame-wise accuracy (in %) of the C3RNN and CRNN methods in estimating the correct number of active sound events in the ANSYN *O*6 subset. The best scores for the respective reference number of overlapping sound events are highlighted. ©2018 IEEE.

	Number of overlapping sound events							
C3RNN	0	1	2	3	4	5	6	Avg.
gcc-ambi	90.7	46.0	38.0	34.4	29.5	10.4	0.0	35.6
mbe-ambi	**92.7**	**66.9**	**56.3**	**47.7**	**34.7**	**16.3**	0.3	**45.0**
mbe-bin	90.4	58.0	46.3	39.8	27.8	12.7	**0.6**	39.4
mbe-mono	89.9	60.8	48.1	35.2	19.1	8.6	0.1	37.4
CRNN								
mbe-ambi	**93.5**	**66.4**	**56.5**	**47.3**	**32.4**	**15.7**	**0.5**	**44.6**
mbe-bin	92.8	60.6	47.7	42.9	29.1	12.3	0.2	40.8
mbe-mono	90.8	59.6	49.9	34.1	18.4	9.7	0.4	37.6

Table 4.7: The frame-wise accuracy (in %) of the C3RNN method in estimating the correct number of active sound events in the TUT-SED 2017 dataset. The best scores for the respective reference number of overlapping sound events are highlighted. ©2018 IEEE.

	Number of overlapping sources				
C3RNN	0	1	2	3	Avg.
mbe-bin	**70.1**	**70.2**	**73.1**	**16.4**	**57.5**
gcc-bin	62.2	64.6	39.4	2.0	42.1

Table 4.8: The SED evaluation scores of proposed C3RNN and baseline CRNN methods for TUT-SED 2017 dataset. The best scores for the respective metric are highlighted. ©2018 IEEE.

	C3RNN		CRNN	
	ER	F	ER	F
mbe-bin	**0.35**	**67.5**	**0.37**	**64.8**
mbe-mono	0.38	64.1	0.39	63.3

filter is convolved with every channel of the input feature unlike in CRNN where it is convolved with just one corresponding input feature channel.

Another observation in Table 4.5 is that the SED performance drops with higher polyphony (*O*3 to *O*6) when using single-channel feature *mbe-mono*, whereas when using multichannel feature *mbe-ambi* there is no such performance drop. Typically, the multichannel feature *mbe-ambi* performs better SED than *mbe-bin*, which in turn performs better than using just single-channel audio feature *mbe-mono*.

Table 4.5 also presents results with *mbe* and *gcc* feature combinations. For the tested cases, the SED performance of using this combination was seen to perform equal to or worse than using just the *mbe* feature. To investigate why this is happening, we carried out an experiment to estimate the number of active sound sources in each frame using just the *mbe-ambi* and *gcc-ambi* features individually. The motivation of using *gcc-ambi* was that it will help recognize overlapping sound events better. From the results in Table 4.6, it seems that the network learns the number of sources per frame better with just the *mbe-ambi* feature. Similar results were obtained with binaural features of ANSYN subsets and TUT-SED 2017 dataset (Table 4.7). Although it was shown that *gcc* feature helped with the TUT-SED 2009 and TUT-SED 2016 datasets in [II], from the results in Table 4.6 and 4.7 it seems that *gcc* is not providing any additional information for the

evaluated datasets. The dominance of *mbe* over *gcc* could be explained by the strong head shadowing effect for different source directions in binaural recordings. Similarly, in the case of multichannel Ambisonic format audio, the spatial direction information is predominantly encoded in inter-channel level differences. But this dominance of *mbe* feature may fail for audio formats which encode the spatial direction information in the phase- or time-differences, and with insignificant level differences. Audio captured with linear microphone arrays and spaced omnidirectional microphones are some examples where the *gcc* feature will provide additional information over *mbe* feature.

Among the different features in Table 4.6 we observe that using multichannel *mbe-ambi* in place of single-channel *mbe-mono* significantly improves the framewise number of active classes estimation. Using *mbe-ambi* improves detection of three overlapping sources by 12.5% and four overlapping sources by 15.6%.

The evaluation results with TUT-SED 2017 dataset is presented in Table 4.8. They are consistent with the results obtained with ANSYN subsets, i.e., the performance of C3RNN and CRNN are comparable. Using binaural *mbe-bin* feature performs better SED than single-channel *mbe-mono* feature.

4.4.3 Contributions and Limitations

In [III] we proposed to use one layer of the 3D convolutional network as the input layer to learn both inter- and intra-channel features jointly. Although the SED performance was comparable to the baseline CRNN method, the proposed C3RNN with 3D convolutional network achieved a lower error rate score in a fewer number of epochs in comparison to CRNN. On the other hand, even though the number of parameters in CRNN and C3RNN is identical, the number of computations per epoch in C3RNN is much higher than CRNN.

We studied the SED performance with an identical polyphonic acoustic scene recorded with single-channel, binaural and multichannel audio. The results obtained were as follows. The required network size increased with the sound scene polyphony. Overlapping sound events are recognized better with multichannel features in comparison to single-channel features. By using multichannel audio instead of single-channel audio, the overall F-score improved by 7.5%, overall ER improved by 10%, and 15.6% more sound events were recognized in time-frames with four overlapping sound events.

5 Sound Event Localization, Detection, and Tracking

In [I], [II] and [III], we showed that the overlapping sound events can be recognized better with multichannel audio. In addition to performing better polyphonic SED, using multichannel recordings also enables the localization and tracking of sound events resulting in MSET output. Further, performing jointly the SED and MSET tasks produces the SELDT output. In this regard, we first discuss the related works for localization, MSET, and SELDT tasks in Section 5.1. Thereafter, we present our proposed methods for localization [IV] in Section 5.2, and SELDT [V, VI] in Section 5.3 and 5.4 respectively. Here, for each publication, we provide the motivation, discuss the method, present the evaluation results and finally provide the summary of our contributions and limitations of the proposed method.

5.1 Related Work

In this thesis, localization refers to estimating the DOA with respect to the microphone. Some popular DOA estimators are based on time-difference-of-arrival (TDOA) [108], steered-response-power (SRP) [109], multiple signal classification (MUSIC) [110] and the estimation of signal parameters via the rotational invariance technique (ESPRIT) [111]. These methods vary in terms of algorithmic complexity, compatibility with different array structures, and model assumptions based on the acoustic scenario. The subspace methods such as MUSIC are generic to array structures and produce high-resolution DOA estimates. However, these subspace methods require a good estimate of the number of active sources to estimate their corresponding DOAs, and this information is not always available. Furthermore, their performance in low SNR and reverberant scenarios is poor [38]. In the rest of this thesis, we refer to these traditional methods as the parametric DOA estimation methods.

To overcome some of the above drawbacks more recent methods have been studying DNN-based DOA estimation. Implementing the DOA using DNN will enable integration of DOA estimation into end-to-end sound analysis and detection systems. In the context of this thesis, the higher-level learning task is SELDT. Table 5.1 presents a brief summary of these methods. All these methods study the localization of point sources that are spatially stationary. Furthermore, these methods were shown to be robust to reverberant scenarios with suitable evaluations. Apart from [104–106] that estimate DOA in a limited azimuth and elevation space, the rest of the methods estimate DOA only along azimuth angle. Although methods [9, 37, 102] studied DOA estimation of up to two temporally overlapping sources, these methods can potentially scale to more sources. Unlike the parametric methods, these methods detect the number of active sources from the input

Table 5.1: Summary of the recent DOA estimation methods using DNNs. 'azi' and 'ele' represents the azimuth and elevation angles of spherical coordinates format, 'x' and 'y' represent the distance along the respective axis of Cartesian coordinates format. If the DOA is estimated in the complete range of the respective format it is denoted as 'Full', otherwise, the estimation is in a limited range. Methods employing regression approach is marked as 'regression', the remaining methods estimate DOAs using classification approach. ©2018 IEEE.

Approach	Input feature	Output format	Sources	DNN	Array	SELDT
Chakrabarty et al. [37, 58]	Phase spectrum	azi	1, multiple	CNN	Linear	×
Yalta et al. [8]	Spectral power	azi (Full)	1	CNN Resnet	Robot	×
Xiao et al. [101]	GCC	azi (Full)	1	FC	Circular	×
Takeda et al. [6, 7]	Eigen vectors of spatial covariance matrix	azi (Full)	1, 2	FC	Robot	×
He et al. [9]	GCC	azi (Full)	Multiple	CNN	Robot	×
Hirvonen [102]	Spectral power	azi (Full) for each class	Multiple	CNN	Circular	×
Yiwere et al. [103]	ILD, cross-correlation	azi and dist	1	FC	Binaural	×
Ferguson et al. [104]	GCC, cepstrogram	azi and dist (regression)	1	CNN	Linear	×
Vesperini et al. [105]	GCC	x and y (regression)	1	FC	Distributed	×
Sun et al. [106]	GCC	azi and ele	1	PNN	Cartesian	×
Roden et al. [107]	ILD, ITD, phase and magnitude spectrum	azi, ele and dist (separate NN)	1	FC	Binaural	×
Proposed [IV]	Phase and magnitude spectrum	azi and ele (Full)	Multiple	CRNN	FOA	×
Proposed [V, VI]	Phase and magnitude spectrum	azi and ele (Full, regression) for each class	Multiple	CRNN	FOA Circular	✓

feature and further estimate their corresponding DOAs. The remaining DNN-based methods algorithmically cannot localize more than one source per time-frame and hence cannot be called a multiple DOA estimator method. All these methods were proposed on different configurations of microphone arrays, thus making it difficult to compare their performance with each other. Finally, these methods use various features such as generalized cross-correlation (GCC), inter-aural level and time differences, or Eigenvectors of the spatial covariance matrix. More recently [37] proposed to use the phase component of the spectrogram as the input feature thereby avoiding any explicit array- or method-specific feature extraction. This was motivated by the fact that the omnidirectional microphones encode the spatial information predominantly in the phase component of the spectrogram.

Early SELDT methods performed SED and DOA estimation separately. This resulted in the data association problem of mapping the multiple recognized sound events with the estimated DOAs [14]. The data association problem is further exacerbated when the estimated number of sound events and DOAs is unequal. Methods such as [68] approached SELDT by recognizing non-overlapping sound events and their corresponding DOA. Since the method estimates only one sound event at a time, the data association problem is not faced. In contrast, [65] proposed the joint localization and recognition of multiple sound events by using a sound-model-based localization and thus overcame the data association problem. Similar joint SED and DOA estimation approaches for multiple sources using DNN were proposed in [102], that mapped the log-spectral power feature to two sound classes in eight azimuth angles using CNN as a multiclass multilabel classification, thus performing the SELDT task. However, all these existing methods focus on the localization of point sources that are spatially stationary and their performance on moving sources is unknown. Further, these methods have only studied localization and detection of sound events and do not perform any explicit tracking. Hence these methods are only performing a sub-task of SELDT, i.e., sound event localization and detection (SELD).

On the other hand, stand-alone MSET methods have been studying tracking of static and moving sources based on spatial information only [112 – 120], additional spectral information [121, 122], or in conjunction with visual information [123]. The general approach of these methods consists of two stages, a frame-wise multiple DOA estimator followed by a temporal tracker such as a Kalman or particle filter. These trackers employ data association algorithms such as Hungarian [56] or Rao-Blackwellized Monte Carlo data association (RBMCDA) [124] to solve the data association problem occurring due to multiple sound events. The MSET methods in the literature are all parametric in nature, and to the best of the author's knowledge, currently, there are no DNN-based methods.

5.2 Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network

At the time of writing [IV], only a limited number of DNN-based DOA estimation methods had shown the ability to detect the number of active sources from the input feature, and further estimate their DOAs. However, none of these methods estimate DOA in complete azimuth and elevation. Thus in [IV], keeping the SELDT task requirements in mind, we proposed to estimate multiple DOAs in complete 2D spherical space. To enable the estimation of DOA in complete 2D spherical space, we employed the FOA format audio. As discussed in Section 2.4.1, FOA is a widely used format for 360° audio representation. Additionally, as an intermediate output, the proposed method produces

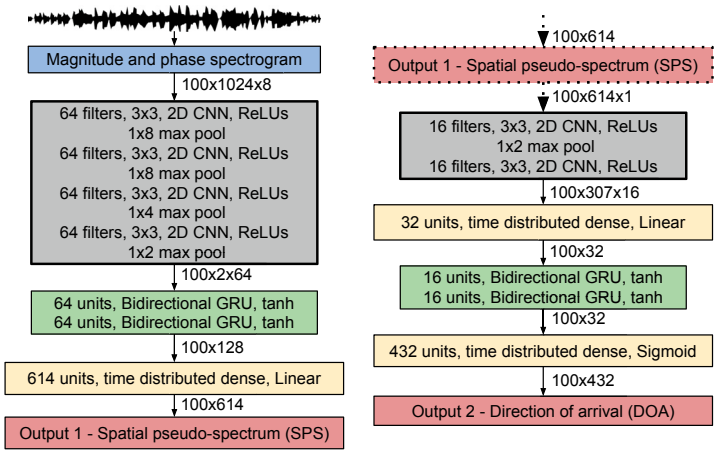


Figure 5.1: DNN structure for direction of arrival estimation of multiple sound sources (DOAnet). ©2018 IEEE.

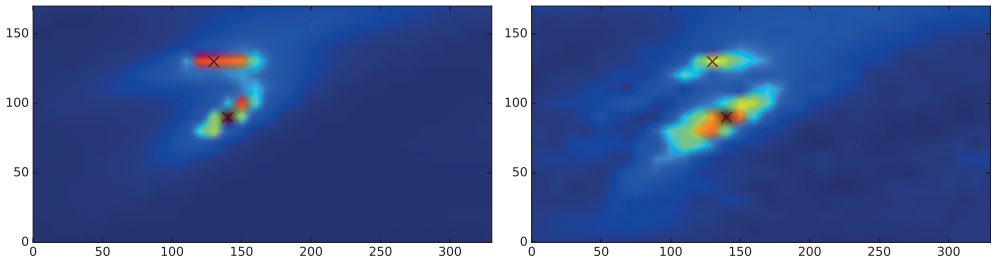
a spatial pseudo-spectrum (SPS), that could be used for soundfield visualizations [125], and room acoustics analysis [126].

5.2.1 Method

The DOA estimation method in [IV] is approached as a multiclass multilabel classification task $P(\mathbf{Y}_{DOA}|\mathbf{X}_{DOA}, \mathbf{W}_{DOA})$. The acoustic model parameters $\mathbf{W}_{DOA}$ are used to map the acoustic features $\mathbf{X}_{DOA}$ to the class-wise probabilities at a fixed U number of azimuth and V number of elevation angles $\mathbf{Y}_{DOA} \in \mathbb{R}^{T\times(U\times V)}$. Additionally, as the intermediate output the input features $\mathbf{X}_{DOA}$ are mapped to SPS $\mathbf{Y}_{SPS} \in \mathbb{R}^{T\times(U\times V)}$ as a regression task $f_{\mathbf{W}_{DOA}}\mathbf{X}_{DOA} :\mapsto \mathbf{Y}_{SPS}$

Figure 5.1 presents the overall structure of the proposed method. As the acoustic features $\mathbf{X}_{DOA}$, the magnitude and phase components of the spectrogram are extracted from the multichannel audio using 2048-point FFT using Hamming windows of 40 ms and 50% overlap. T frames of 1024 spectral values corresponding to the positive frequencies without the zeroth bin for each of the K channels resulting in a $T \times 1024 \times 2K$ as the output of the feature extraction block. The $2K$ represents the magnitude and phase components for each channel. The $T \times 1024 \times 2K$ dimension sequence is first mapped to an SPS using a CRNN as a multioutput regression task. Further, the SPS is mapped to DOA output using a second CRNN as a multiclass multilabel task. The two CRNNs are trained jointly and together form the acoustic model $\mathbf{W}_{DOA}$, referred to as DOAnet hereafter.

The DOA output $\mathbf{Y}_{DOA} \in \mathbb{R}^{T\times(U\times V)}$ is of the dimension $T \times 432$, where $432 = 36 \times 12 = U \times V$. The number of azimuths U and elevation angles V is derived from the dataset ANSYN and RESYN (see Section 2.4.1) used for evaluation. These datasets contain sound events spatially located at 10° resolution along both azimuth and elevation, with the elevation angle $\theta \in [-60, 60)$. During inference, the probability at these 432 DOA coordinates are thresholded with a value of 0.5, anything greater suggests the presence of the source in the respective coordinate otherwise their absence. Similarly, for the SPS output $\mathbf{Y}_{SPS} \in \mathbb{R}^{T\times(U\times V)}$ of DOAnet, although the same 432 coordinates could be used,



(a) MUSIC estimated (b) DOAnet estimated

Figure 5.2: SPS visualization of two closely located sources for the baseline MUSIC and proposed DOAnet method. The reference position of the DOA is represented with the black-cross mark. The horizontal and vertical axes represent the azimuth and elevation angles in degrees. ©2018 IEEE.

we extend the elevation angle to the full range $\theta \in [-90, 90]$ and obtain SPS of dimension $T \times 614$.

During DOAnet training, the DOA labels are obtained from the reference of the datasets, whereas the SPS labels are obtained by using the MUSIC algorithm [110] to generate SPS at the same frame rate as the DOAnet. The DOAnet is trained for 1000 epochs using the Adam optimizer with MSE loss for the SPS output, and cross-entropy loss for the DOA output. An equal weighted sum, of the two losses was used for backpropagation. Early stopping was used to halt the training if the MSET score (see Section 2.4.2) did not improve for 100 epochs.

5.2.2 Evaluation

The DOAnet was evaluated on ANSYN and RESYN datasets. The hyperparameters were tuned with respect to the ANSYN $O1$ subset and the configuration obtaining the best DOA metric score is as shown in Figure 5.1. This configuration has 677 K weights, and the same configuration is used for the remaining subsets. The DOA estimation performance of DOAnet was compared with the MUSIC algorithm [110]. MUSIC is a popular high-resolution DOA estimation method that can be applied to generic array

Table 5.2: The evaluation scores of DOAnet for different datasets, and the corresponding results with baseline MUSIC. ©2018 IEEE.

	ANSYN			RESYN Room 1		
Max. no. of overlapping sources	1	2	3	1	2	3
SPS SNR (in dB)	9.90	3.35	-0.26	3.11	1.24	0.13
DOAnet with unknown number of active sources (threshold of 0.5)						
DOA error	0.57	8.03	18.34	6.31	11.46	38.41
Frame recall	95.4	42.7	1.8	59.3	15.8	1.2
DOA error with known number of active sources						
DOAnet	1.14	27.52	49.30	12.61	38.98	67.07
MUSIC	2.29	8.60	28.66	25.80	57.33	91.72

Table 5.3: The evaluation scores of DOAnet and MUSIC for unmatched reverberant room. ©2018 IEEE.

	RESYN Room 2		RESYN Room 3	
Max. no. of overlapping sources	1	2	1	2
SPS SNR (in dB)	3.53	1.49	3.49	1.46
DOAnet error (Unknown number of sources)				
DOA error	3.44	6.88	4.59	10.89
Frame recall	46.2	14.3	49.7	14.1
DOA error (Known number of sources)				
DOAnet	8.60	32.10	9.17	33.82
MUSIC	31.52	58.47	33.25	60.76

setups. MUSIC produces similar SPS output and can detect multiple DOA estimates in complete azimuth and elevation space. Hence the SPS and DOA outputs of the proposed DOAnet can be directly compared with the corresponding MUSIC output. The MUSIC algorithm requires the spectrogram, and the knowledge of the number of active sources in each time-frame as the input to estimate their respective DOAs and SPS. The number of active sources knowledge is obtained from the reference of the studied datasets. During the evaluation, we use a similar spatial grid as DOAnet for MUSIC, i.e., complete azimuth and elevation space at 10° resolution.

The SPS output of the DOAnet and baseline MUSIC when two active sources are closely located is visualized in Figure 5.2. We can observe that the two peaks are well separated in the DOAnet SPS, similarly to the MUSIC estimate. Further, the SPS estimates of DOAnet is quantitatively compared against MUSIC using the SNR metric (see Section 2.4.2) and the results are presented in Table 5.2. The high SNRs for $O1$ and $O2$ subsets of the respective datasets show that the SPS estimated by DOAnet is comparable to MUSIC SPS. In the case of $O3$ subsets, the MUSIC is already at its theoretical limit of estimating $N-1$ sources from N-dimensional signal space [30], i.e., MUSIC is estimating SPS and DOAs for three ($O3$) sources from a four-channel FOA signal, which will result in weak SPS and DOA estimates. Thus training DOAnet on this weak MUSIC SPS estimate results in poor performance. Ideally, with more number of channels, which the proposed DOAnet can easily extend to, the DOAnet can potentially estimate more than two sources.

(a) ANSYN $O1$ (b) ANSYN $O2$ (c) RESYN $O1$ (d) RESYN $O2$

Figure 5.3: The performance of the DOAnet in estimating the number of active sources in each frame visualized as a confusion matrix. The horizontal axis represents the estimate of DOAnet, and the reference is represented along the vertical axis. ©2018 IEEE.

The DOA error of DOAnet obtained without the knowledge of the number of active sources is considerably better than baseline MUSIC that requires the number of sources knowledge as seen in Table 5.2. Especially in the case of RESYN dataset, this difference is significant. However, the number of estimated sound events by DOAnet are few. For example in ANSYN $O2$ subset the DOAnet estimated the correct number of sound events in only 42.7% of the frames. Figure 5.3, presents the confusion matrix of the number of estimated DOAs for the studied datasets. In general, the DOAnet performance is seen to drop for a higher number of overlapping sound events.

Since MUSIC uses the knowledge of the number of active sources, we did a study of using the same knowledge for DOAnet. We use this knowledge to choose the number of DOAs from the prediction layer of DOAnet without thresholding it with a value of 0.5. With this knowledge, the DOA error of the DOAnet continues to be better in the reverberant context, but MUSIC continues to perform better in the anechoic context (Table 5.2). This is still a good result considering the simple strategy for choosing the DOAs in DOAnet.

The result of training the DOAnet on the RESYN Room 1 dataset, and testing it on the unmatched reverberant Room 2 and Room 3 of the RESYN dataset is presented in Table 5.3. The results were observed to be consistent with the performance in Room 1 data in Table 5.2. This is a significant result, enabling us to use a model trained on one reverberant scene, on other moderately mismatched reverberant scenes.

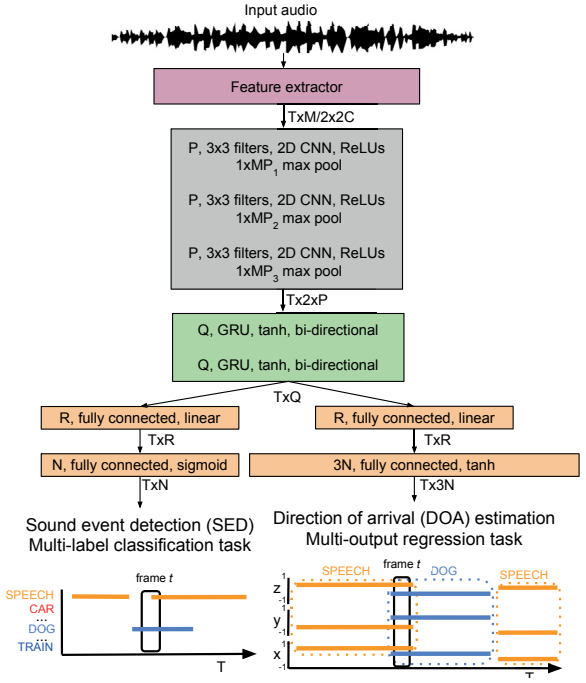
5.2.3 Contributions and Limitations

In [IV] we presented a DNN-based method for DOA estimation of multiple temporally overlapping sound events in complete azimuth and elevation space that produces spatial pseudo-spectrum. This is the first DNN-based method estimating multiple DOAs in complete azimuth and elevation space, and producing a pseudo-spectrum. The proposed DOAnet is shown to learn the number of active sources from the input acoustic feature, and further estimate their respective DOAs. Unlike most of the other recent methods presented in Table 5.1 that employ method- or array-specific acoustic features, the proposed DOAnet uses generic low-level phase and magnitude spectrogram as features. The DOAnet is shown to localize the DOAs better than the parametric MUSIC algorithm which requires knowledge of the number of active sources. The difference in performance between DOAnet and MUSIC is increased for reverberant scenarios. It was also shown that the DOAnet trained on one reverberant scene can be used in moderately mismatched reverberant scenes without a drop in performance.

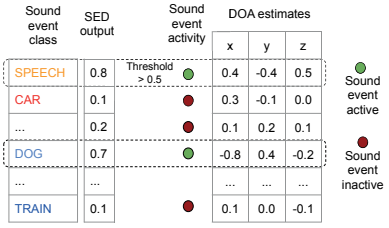
However, it was observed that the number of DOAs estimated by the DOAnet reduces with the higher number of overlapping sound events. Further, since MUSIC was used as a baseline, the DOAnet was trained using MUSIC SPS to have a fair comparison. Ideally, given a dataset with reference DOA locations, the SPS can be generated directly from the DOA and used to train the DOAnet. Alternatively, the DOAnet can be trained without the SPS to generate DOAs directly. Further, the IRs employed for the RESYN datasets were synthesized using the image source method, which does not completely reflect the properties of a real, measured IR such as scattering, diffuse reflections, and measurement errors. As a result, synthetic IRs can be easier to learn for DNN-based methods. In the future, similar experiments with measured IRs have to be carried out to validate the performance of DOAnet in real-life acoustic scene.

5.3 Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Network

Although the DOAnet proposed in [IV] produced better DOA estimates than the parametric MUSIC algorithm, the frame recall of DOAnet was poor for a higher number of temporally overlapping sound events. To overcome this, in [V] we proposed to use the state-of-the-art SED output from the CRNN in [II] to estimate individual class-wise DOAs when the sound event was active thus performing SELDT. Further, the DOAnet was only evaluated on reverberant dataset generated using synthetic IRs and the performance on measured IRs was unknown. Hence, in this section, we investigate the performance of the proposed SELDT method on separate datasets that were generated with synthetic and measured IRs. Finally, in this section, we only investigate the SELDT performance of the proposed method for spatially stationary sources. The performance for spatially moving sources is investigated in Section 5.4.



(a) SELDnet



(b) SELDnet output

Figure 5.4: a) The proposed method for joint localization and detection of sound events (SELDnet). b) The output of SELDnet for an individual frame t shown in a). A sound event is considered to be localized and detected when the SED output probability exceeds the threshold. ©2018 IEEE.

5.3.1 Method

In the proposed method, the SED is approached as a multiclass multilabel classification task $P(\mathbf{Y}_{SED}|\mathbf{X}, \mathbf{W}_{SELDT})$, whereas the MSET is approached as multioutput regression task $f_{\mathbf{W}_{SELDT}} : \mathbf{X} \mapsto \mathbf{Y}_{MSET}$. The classwise probabilities of SED $\mathbf{Y}_{SED} \in \mathbb{R}^{T \times C}$ and the

continuous MSET output $\mathbf{Y}_{MSET} \in \mathbb{R}^{T \times 2GC}$ are both obtained from the input features $\mathbf{X}$ using the same acoustic model $\mathbf{W}_{SELDT}$.

Figure 5.4 shows the proposed SELDT method. As the input feature $\mathbf{X}$ we use the phase and magnitude component of the spectrogram extracted from each of the K channels using an M-point FFT with a Hamming window of length M and 50% overlap. The feature extraction step produces a feature of dimension $T \times M/2 \times 2K$, where T is the number of time frames with spectral features corresponding to the $M/2$ positive frequencies. The $2K$ represents the phase and magnitude components of the FFT for the K channels. As the acoustic model $\mathbf{W}_{SELDT}$ we use a CRNN that maps the input feature sequence to two outputs - SED $\mathbf{Y}_{SED}$ and MSET $\mathbf{Y}_{MSET}$. At the first output, for a T frame input sequence the SED results in a $T \times C$ dimension output corresponding to the temporal activity for each of the C sound event classes in the dataset. At the second output, we estimate just one DOA instance ($G = 1$) for each of the C sound event classes using 3D Cartesian coordinates of DOA on a unit sphere, hence resulting in a $T \times 3C$ dimension output.

The SED outputs correspond to class-wise probabilities $\in [0, 1]$ as shown in Figure 5.4b. These probabilities are thresholded with a value of 0.5 to obtain the binary decision of class activity. Finally, the DOA for the active classes is obtained by choosing the corresponding three regressors estimating the DOA. The SED classification layer employs the sigmoid activation which enables multilabel multiclass classification. The DOA estimating regressors employ the tanh activation whose range $\in [-1, 1]$ corresponds to the extent of the unit sphere in the respective axes. This architecture is referred to as SELDnet hereafter and is trained with cross-entropy loss for SED output and MSE loss for MSET. A weighted sum of the two losses is used for backpropagation. We use the Adam optimizer to train SELDnet for 1000 epochs and use early stopping to halt the training if the SELDT score does not improve for 100 epochs.

5.3.2 Evaluation

The SELDnet is evaluated with seven synthesized datasets consisting of spatially stationary sources that are listed in Table 2.1 under SELDT - Static sources. These datasets represent different acoustic scenarios such as anechoic and reverberant; different array configurations - Ambisonic and circular array; different polyphony - up to one, two and three temporally overlapping sound events; simulated with different impulse responses - synthetic and real-life and different dataset sizes.

Table 5.4: Summary of the baseline and proposed methods for SED, DOA estimation and SELD tasks. ©2018 IEEE.

Task	Acronym	Notes	Datasets evaluated
SED	SEDnet [II]	Single channel	All
	MSEDnet [II]	Multichannel	
DOA	MUSIC*[110]	Azi and ele	All except CANSYN and CRESYN
	DOAnet [IV]	Azi and ele	
	AZInet [37]	Azi	CANSYN and CRESYN
SELD	HIRnet [102]	Azi	
	SELDnet-azi	Azi	All
	SELDnet	Azi and ele	

*Parametric, all other methods are DNN-based

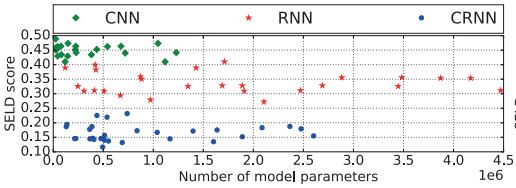


Figure 5.5: The SELDT scores obtained with different configurations of CNN, RNN and CRNN architecture for ANSYN *O*2 dataset. ©2018 IEEE.

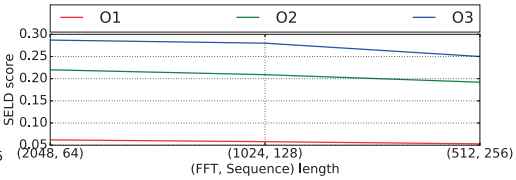


Figure 5.6: The SELDT scores obtained with different combinations of FFT points and input sequence length (in frames) for ANSYN subsets. ©2018 IEEE.

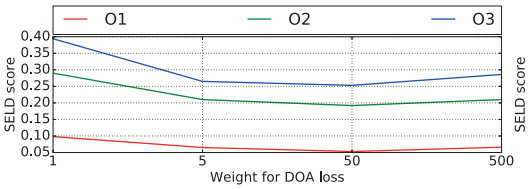


Figure 5.7: The SELDT scores obtained with different weights of DOA output for ANSYN subsets. ©2018 IEEE.

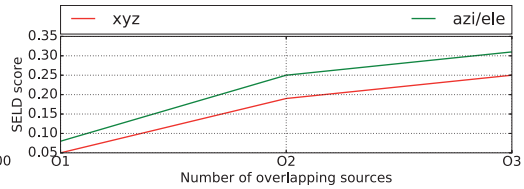


Figure 5.8: The SELDT score obtained with different DOA output formats for ANSYN subsets. ©2018 IEEE.

The SELDnet performance is compared to six different baselines as summarized in Table 5.4. This includes two SED baselines (single- and multichannel), three DOA baselines (parametric and DNN-based), and a SELD baseline. The SED baselines are the single-channel (SEDnet) and multichannel (MSEDnet) versions of CRNN proposed in [II]. Unlike in [II], particularly for this study, these methods are trained with the spectrogram instead of log mel-band energies. This was done to have a direct comparison with the SELDnet that uses the spectrogram as an input feature.

As the parametric baseline for DOA estimation we use MUSIC [110], and as the DNN baselines, we use DOAnet [IV] and AZInet [37]. AZInet is a CNN-based method that approaches DOA estimation along azimuth as a multiclass multilabel classification. Since AZInet was proposed for omnidirectional microphone arrays, it employs just the phase component of the spectrogram as the acoustic feature. To have a direct comparison with AZInet, we change the SELDnet output to estimate DOA along x, y axes only. This version of SELDnet is referred to as SELDnet-azi hereafter and is evaluated along with AZInet on the circular array datasets CANSYN and CRESYN. The tracking capability of SELDnet is evaluated with suitable baselines in Section 5.4.

Similar to AZInet, the SELD baseline HIRnet [102] is a CNN-based method that maps the log-spectral power of each channel to sound event activity and their respective azimuth locations using multiclass multilabel classification. Since the HIRnet was proposed for omnidirectional microphones we evaluate its performance along with SELDnet-azi on CANSYN and CRESYN datasets.

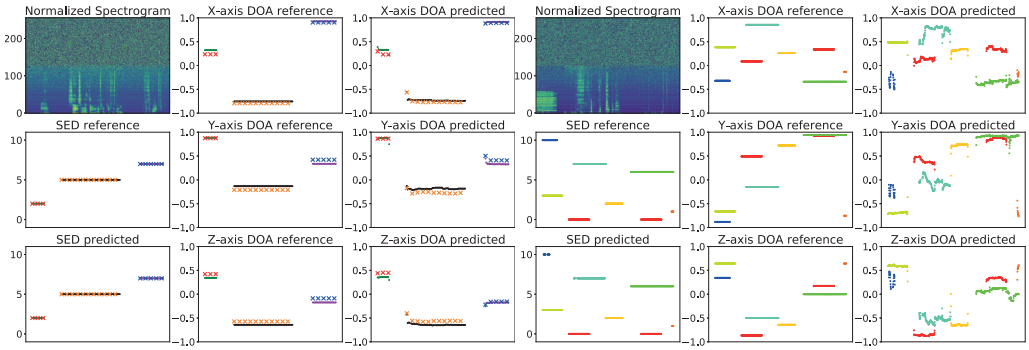
A wide variety of architectures, including CNN, RNN and FC individually and as combinations were explored with the ANSYN *O*2 subset with frame length $M = 1024$ ($23.2ms$). Within each architecture different configurations including the number of layers and number of units per layer were varied. The SELDT score achieved corresponding to the number of parameters in these architectures are visualized in Figure 5.5. The results

show that the CRNN architecture, in general, performs better than CNN and RNNs individually. On performing hyperparameter tuning of CRNN, the optimum parameter across ANSYN subsets had three convolutional layers with $P = 64$ (in Figure 5.4a) filters, followed by two recurrent layers with $Q = 128$ GRU, and one FC layer each for SED and DOA estimation with $R = 128$ units. A max pooling $MP_i = (8, 8, 2)$ for the three convolutional layers was seen to give the best results. This configuration had 513 K parameters. As SELDnet has two outputs, we tuned the weights for the weighted combination during backpropagation. It was observed that weighting the DOA output 50 times more than SED gave the best results as shown in Figure 5.7. Weighting the SED output more than DOA gave poorer results. A sequence length of $T = 512$ ($2.97s$) and $M = 512$ gave the best results across ANSYN subsets. The same configuration with a reduced sequence length of $T = 256$ gave the best results for RESYN, CANSYN and CRESYN datasets. Model parameters identical to ANSYN gave the best performance for REAL, REALBIG and REALBIGAMB subsets.

The regression-based DOA estimation can output either the azimuth and elevation angle in spherical coordinates or the Cartesian coordinates of x, y, z. To choose between the two formats, we studied their respective SELDT performance. During this study, we use the default azimuth angle of 180° and elevation of 60° when the sound event is not active. Similarly for the Cartesian coordinates, we use $x = 0$, $y = 0$, and $z = 0$. The chosen Cartesian coordinates are equidistant from all the possible DOA values, whereas no such default value exists for the spherical coordinates. Hence a default value in the same order as the respective azimuth and elevation angles are chosen as their default values. Additionally, the activation of the DOA output was changed to linear from tanh for spherical coordinates to support the range of angles. From Figure 5.8 we see that the SELDT performance improves by using the Cartesian coordinates instead of spherical coordinates. This suggests that the discontinuity around azimuth angle of 180° in spherical coordinates which do not arise in the Cartesian coordinates is reducing the overall DOA estimation performance.

Theoretically, using regression instead of classification for DOA estimation enables continuous DOA estimation, as long as the model learns the correct mapping. With the classification approach, we are limited to the fixed set of angles the model is trained on, and it cannot scale to new angles unless retrained with it. We study the performance of SELDnet for such unseen angles by generating a test set of ANSYN $O1$ that is identical in terms of the temporal sound event location, but the spatial position of the sound events are shifted by 5° along both azimuth and elevation. Since all the sound events in the training split were synthesized at a resolution of 10°, this shift of 5° makes the DOA location unseen. The results of SELDnet trained on ANSYN and tested on this shifted data can be visualized in Figure 5.9a. All the subplots visualize a 1000 frame sequence, and each sound event class is represented with a unique color. The bold lines represent the SELDT results on the original unshifted test data, whereas the × marker represents the SELDT results on the DOA shifted test data. We see that the SED performance is identical, and the DOA values are shifted accordingly. This shows that the regression-based DOA approach truly helps in learning continuous mapping of the spatial location, and works seamlessly on unseen DOA locations.

Figure 5.9b shows a similar visualization of the SELDnet input and output for the ANSYN $O2$ subset. The SED performance is observed to be accurate, while the DOA predictions are seen to be varying around the respective mean reference value. We believe this is a result of our training procedure. This can actually be observed Figure 5.9b, in the



(a) ANSYN $O1$ **(b)** ANSYN $O2$

Figure 5.9: Visualizations of the input and outputs of SELDnet for ANSYN $O1$ and $O2$ datasets. Across the sub-plots, the horizontal axis represents the same time frames. The vertical axis of the spectrogram sub-plot represents the frequency bins, the SED sub-plots represent the unique sound event class identifier, the DOA sub-plots represent the distance from the origin along the respective axis. The bold line is used to represent both the reference and SELDnet predictions for test data of ANSYN $O1$ and $O2$ subsets. The $\times$ marker represents the results of the SELDnet on ANSYN $O1$ test data with *unseen* DOA values that are shifted by 5° in both azimuth and elevation. The spectrogram sub-plot visualizes both the magnitude and phase components. ©2018 IEEE.

spectrogram. Even though the individual sound event instances have inherent silences and amplitude modulations, the reference label for SED and DOA is constant and does not include this information. From the SELDnet predictions, we observe that these variations have no effect on the SED output but leads to fluctuating DOA estimates. From these visualizations, we see that the proposed SELDnet successfully estimates the number of active sources, identifies their temporal activity, recognizes their sound class and further localizes and tracks them simultaneously.

The quantitative performance of SELDnet on ANSYN subsets is presented in Table 5.5. Among the SED baselines, we observe that the proposed SELDnet achieves SED performance comparable to the best SED baseline MSEDnet. With respect to DOA metrics, the frame recall of SELDnet has improved significantly in comparison to DOAnet at the expense of a slight increase in DOA error. However, for $O2$ and $O3$, MUSIC, which relies on knowledge of the number of active sources, achieves the lowest DOA error.

Parametric DOA estimators like MUSIC are sensitive to reverberation. To compare the performance of SELDnet in a reverberant scenario, we evaluated on the RESYN, REAL, REALBIG, and REALBIGAMB subsets. For practical or commercial applications, it may not be feasible to train SELDnet for every reverberant scenario (room dimension, surface material distribution, and reverberation). Ideally, if the SELDnet is robust to a moderate mismatch in reverberant scenarios, then a single model can be used in a range of comparable room configurations. To study this, we train the SELDnet on the RESYN Room 1 split and test it on the RESYN Room 2 and Room 3 splits (see Section 2.4.1). From the results in Table 5.5, for RESYN Room 1, the MUSIC performs poorly in comparison to both DOAnet and SELDnet with respect to the DOA error. SELDnet significantly outperforms DOAnet in terms of the frame recall. This shows that the SELDnet is robust to reverberant scenarios. With regard to mismatched reverberant scenes, the performance of SELDnet trained on RESYN Room 1 remains consistent in

Table 5.5: The evaluation scores of the proposed SELDnet and respective baselines for ANSYN and RESYN Room 1 datasets. The RESYN Room 2 and 3 results were obtained from SELDnet trained on Room 1. The best scores for the respective subsets are highlighted. ©2018 IEEE.

		ANSYN			RESYN Room 1			RESYN Room 2			RESYN Room 3		
	Overlap	1	2	3	1	2	3	1	2	3	1	2	3
SED metrics													
SELDnet	ER	**0.04**	0.16	0.19	**0.10**	0.29	0.32	**0.11**	0.33	0.35	**0.13**	0.32	0.34
	F	**97.7**	89.0	85.6	**92.5**	**79.6**	**76.5**	**91.6**	**79.5**	**75.8**	**89.8**	79.1	75.5
MSEDnet [II]	ER	0.10	**0.13**	**0.17**	0.17	**0.28**	**0.29**	0.19	**0.30**	**0.26**	0.18	**0.29**	**0.30**
	F	94.4	**90.1**	**87.2**	89.1	79.1	75.6	88.3	78.2	74.2	86.5	**80.5**	**76.1**
SEDnet [II]	ER	0.14	0.16	0.18	0.18	**0.28**	0.30	0.19	0.32	0.28	0.21	0.32	0.33
	F	91.9	89.1	86.7	88.2	76.9	74.1	87.6	76.4	73.2	85.1	78.2	75.6
MSET metrics													
SELDnet	DOA error	3.4	13.8	17.3	9.2	20.2	**26.0**	11.5	26.0	**33.1**	12.1	25.4	**31.9**
	Frame recall	**99.4**	**85.6**	**70.2**	**95.8**	**74.9**	**56.4**	**96.2**	**78.9**	**61.2**	**95.9**	**78.2**	**60.7**
DOAnet [IV]	DOA error	**0.6**	8.0	18.3	**6.3**	**11.5**	38.4	**3.4**	**6.9**	-	**4.6**	**10.9**	-
	Frame recall	95.4	42.7	1.8	59.3	15.8	1.2	46.2	14.3	-	49.7	14.1	-
MUSIC	DOA error	4.1	**7.2**	**15.8**	40.2	47.1	50.5	45.7	58.1	74.0	48.3	60.6	75.6

Table 5.6: The evaluation scores of the proposed SELDnet and respective baselines for REAL, REALBIG, and REALBIGAMB datasets. The best scores for the respective subsets are highlighted. ©2018 IEEE.

		REAL			REALBIG			REALBIGAMB 20dB			REALBIGAMB 10dB			REALBIGAMB 0dB		
	Overlap	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
SED metrics																
SELDnet	ER	0.40	0.49	0.53	0.37	0.42	0.50	**0.34**	0.46	0.52	**0.37**	0.49	0.52	**0.46**	0.58	0.59
	F	60.3	53.1	51.1	65.4	61.5	56.5	65.6	58.5	55.0	**66.3**	55.4	53.3	**57.9**	48.6	49.0
MSEDnet [II]	ER	**0.35**	**0.38**	**0.41**	**0.34**	**0.39**	**0.38**	0.35	**0.40**	**0.41**	0.38	**0.43**	**0.42**	0.48	**0.56**	**0.54**
	F	**66.2**	**61.6**	**59.5**	67.3	61.8	61.9	**66.0**	**61.6**	**60.1**	63.2	**58.7**	**59.3**	54.5	**49.3**	**51.3**
SEDnet [II]	ER	0.38	0.42	0.43	0.38	0.43	0.44	0.39	0.42	0.43	0.41	0.44	0.46	0.51	0.61	0.57
	F	64.6	61.5	57.2	**68.0**	**62.4**	**62.4**	65.7	60.1	59.2	62.7	56.3	56.9	52.6	46.0	50.4
MSET metrics																
SELDnet	DOA error	26.6	33.7	36.1	23.1	31.3	34.9	25.4	32.5	36.1	27.2	32.5	36.1	30.7	33.7	36.7
	Frame recall	**64.9**	**41.5**	**24.6**	**68.0**	**45.2**	**28.3**	**69.1**	**42.8**	**25.8**	**66.9**	**40.0**	**27.3**	**62.5**	**35.2**	**23.4**
DOAnet [IV]	DOA error	**6.3**	**20.1**	**25.8**	**7.5**	**17.8**	**22.9**	**6.3**	**18.9**	**25.78**	**8.0**	**20.1**	**24.1**	**14.3**	**24.1**	**27.5**
	Frame recall	46.5	11.5	2.9	44.1	12.5	3.1	34.7	11.6	3.2	42.1	13.5	3.3	30.1	10.5	2.8
MUSIC	DOA error	36.3	49.5	54.3	35.8	49.6	53.8	54.5	56.1	61.3	51.6	54.5	62.6	41.9	47.5	62.3

Table 5.7: The evaluation scores of the proposed SELDnet and respective baselines for CANSYN and CRESYN datasets. The best scores for the respective subsets are highlighted. ©2018 IEEE.

		CANSYN			CRESYN		
	Overlap	1	2	3	1	2	3
SED metrics							
SELDnet	ER	0.11	**0.18**	0.19	0.13	0.22	0.30
	F score	93.0	86.6	85.3	90.4	82.2	78.0
SELDnet-azi	ER	**0.08**	0.19	0.24	**0.06**	**0.18**	**0.20**
	F score	**94.7**	87.5	83.8	**96.3**	**87.9**	**85.6**
MSEDnet [II]	ER	0.09	**0.18**	**0.16**	0.12	0.22	0.26
	F score	94.6	**89.0**	**86.7**	92.7	83.7	80.7
SEDnet [II]	ER	0.15	0.21	0.20	0.18	0.26	0.25
	F score	91.4	87.3	84.7	90.5	84.3	82.8
HIRnet [102]	ER	0.41	0.45	0.62	0.43	0.46	0.50
	F score	60.0	54.9	58.8	59.3	60.2	58.6
DOA metrics							
SELDnet	DOA error	29.5	31.3	34.3	28.4	33.7	41.0
	Frame recall	97.9	78.8	67.0	96.4	75.7	60.7
SELDnet-azi	DOA error	7.5	14.4	19.6	5.2	13.2	18.4
	Frame recall	98.0	**82.1**	**66.2**	**98.5**	**82.3**	**70.6**
HIRnet [102]	DOA error	5.2	16.3	33.0	7.4	18.6	43.3
	Frame recall	60.2	35.9	18.4	56.9	20.5	10.7
AZInet [37]	DOA error	**1.2**	**4.0**	7.4	**2.3**	**6.9**	**9.7**
	Frame recall	**99.4**	80.5	60.5	97.3	65.2	44.8

Room 2 and 3. This is a significant result, allowing a SELDnet trained in one reverberant scene to be used in comparable reverberant scenarios.

For the real-life IR datasets REAL, the overall performance of SELDnet deteriorated in comparison to the ANSYN and RESYN datasets, as seen in Table 5.6. A similar drop in performance with real-life datasets has also been reported in SED study [61]. In general, the frame recall of SELDnet is significantly better than DOAnet, whereas the DOA error of DOAnet is better than SELDnet. With regard to SED metrics, the baseline MSEDnet is observed to perform the best. With the larger real-life dataset REALBIG, the SELDnet performance was seen to improve both the SED and DOA estimation. In comparison, the margin of improvement for the SED baselines was smaller than SELDnet. A similar study was carried out with larger simulated datasets ANSYN and RESYN, but the SELDnet performance did not show much improvement. This suggests that performing SELDT task on real-life datasets is more challenging than simulated datasets, hence larger real-life datasets are required for better performance of deep learning models.

The performance of SELDnet was seen to be consistentfor signal-to-noise ratios (SNRs) of 10 and 20 dB, as seen in Table 5.6, but was observed to drop for an SNR of 0 dB. Overall, the trend of better frame recall and poor DOA error in comparison to DOAnet continued with these datasets as well. The parametric MUSIC algorithm was seen to perform poorly across the real-life datasets.

SELDnet is a generic method that can learn to recognize the sound events and their respective spatial locations from any microphone array configuration. To prove this, the SELDnet is evaluated on circular-array datasets CANSYN and CRESYN. Unlike the Ambisonic datasets studied so far, the circular-array studied has a different number of microphones, all located in a single plane, and with an omnidirectional response. Table 5.7 presents the results of SELDnet on circular-array datasets. Among SED

performances, the SELDnet-azi is observed to perform the best among most subsets followed by the MSEDnet. SELDnet-azi is also seen to perform the best in terms of frame recall, while AZInet performed the best in terms of DOA error. Between the SELDnet and SELDnet-azi, even though the frame recall is comparable, the DOA error of SELDnet is poorer, suggesting that the localization in 2D is difficult with a circular array using the proposed SELDnet architecture.. In comparison to the SELDnet performance on Ambisonic datasets, the results on circular array datasets are comparable, suggesting that the SELDnet can learn similar information independent of the microphone array.

5.3.3 Contributions and Limitations

In [V], we presented the first SELDT method that can localize and track multiple overlapping sound events in complete azimuth and elevation space. In comparison to the only previous DNN-based SELD baseline [102], the proposed method performed better SED and DOA estimation. By using the SED output as the confidence measure, SELDnet achieved a higher frame recall of DOAs in comparison to baselines and our previous DOAnet [IV]. The SELDnet was seen to be robust to mismatched reverberant scenarios, allowing the use of a trained model on comparable reverberant scenarios. Finally, we showed that the proposed SELDnet is a generic approach independent of the input microphone array used. To support open research and reproducibility, all the datasets used and the SELDnet method have been made publicly available[1]. Further, to increase the visibility and make it easier for researchers to explore the SELDT task, a research challenge was organized at DCASE 2019[2].

The MSET output of SELDnet is performed using regression that allows estimation of unseen DOA values unlike the classification approach in [37][IV] that is limited to a fixed number of spatial locations. By using the regression-based MSET instead of a classification-based approach we overcome the imbalanced and large number of output classes problems discussed in 2.3.

Although the SELDnet achieved a high frame recall of DOAs, the corresponding DOA error was poorer compared to baselines estimating the DOA as multiclass multilabel classification across datasets. We believe this is a result of the regression-based DOA estimation in SELDnet not having learned the DOA mapping between input feature and corresponding DOAs completely. This can potentially be improved with larger datasets and more powerful models. Furthermore, by estimating only one location for each sound class, the SELDnet model overlooks the scenes where a single class can occur simultaneously in multiple different locations. A potential approach to overcome this with the current architecture might be to estimate multiple DOAs for each sound event class. In general, the choice of classification only [102], or classification-regression-based SELDT approach like SELDnet can be made based on the required frame recall, DOA error, DOA resolution, robustness to unseen DOA values, robustness to mismatched scenarios, and development dataset size.

[1] https://github.com/sharathadavanne/seld-net
[2] http://dcase.community/challenge2019/task-sound-event-localization-and-detection

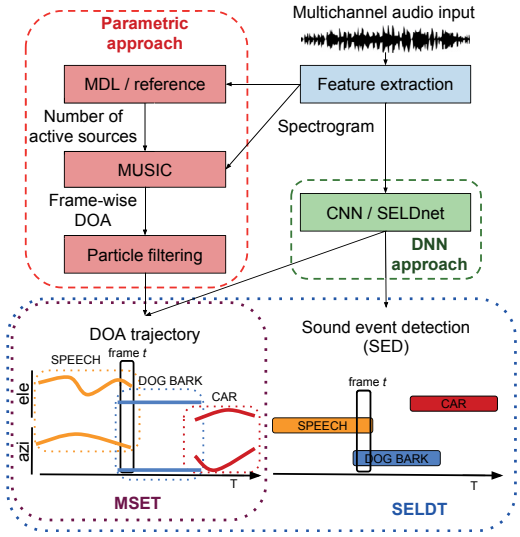


Figure 5.10: The workflow of DNN-based approach producing SELDT results and the parametric baseline approach producing MSET results. The coloring and naming of sound classes in the MSET output have been done only to visualize the concept better. MSET results in real-life as seen in Figure 5.12 do not contain the sound class information.

5.4 Localization, Detection, and Tracking of Multiple Moving Sources with Convolutional Recurrent Neural Network

The SELDnet proposed in [V] was evaluated on static scenes with sound events that are spatially stationary. In a real-life acoustic scene, the sound events are not always stationary but can move around with varying speeds. The localization, detection, and tracking performance of SELDnet in such a realistic sound scene is evaluated in [VI] with suitable datasets.

We show that the recurrent layers of the SELDnet are the key to successful tracking performance. Since this is the first DNN-based method performing tracking, we compare the tracking performance with a parametric MSET method combining MUSIC for frame-wise multiple DOA estimation, and an RBMCDA particle filter [124] for estimating the onset, offset, and temporal trajectories for individual sound events.

5.4.1 Method

The workflow for the studied SELDnet and the parametric MSET methods are shown in Figure 5.10. The input to the two methods is the spectrogram of dimension $T \times F \times K$, where F is the number of positive frequency bins and K is the number of microphone channels. The spectrogram is obtained for each of the K channels using a 2F-point discrete Fourier transform with a Hamming window of length $2F$ and 50 % overlap.

As discussed in Section 3.5 the current output of a recurrent layer is influenced by both the current input and the input from the previous frames. This process is similar to parametric MSET methods such as particle filters, which use both the input at the current time-frame and the knowledge accumulated from the previous time-frames to predict the output for the current time-frame. A more theoretical relation between particle filters

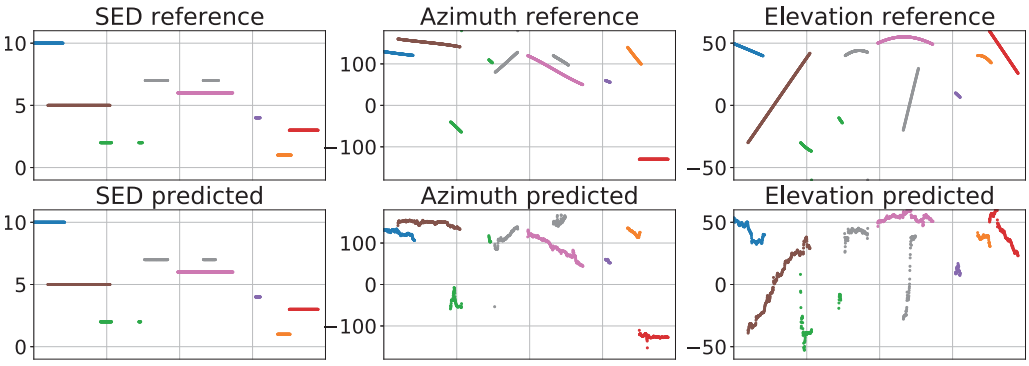


Figure 5.11: Visualization of the reference and SELDnet predictions for moving sources in MANSYN $O2$ subset. The horizontal axis represents the same time-frames in all sub-plots. The vertical axis in SED sub-plots represents unique sound event class identifier and DOA azimuth and elevation angles for remaining subplots.

and recurrent layers is established in [127]. Thus, by jointly training the shared recurrent layers with SED and MSET loss in SELDnet, the recurrent layers learn to associate DOAs from neighboring frames belonging to the same class. This association produces SELDT results. In comparison to MSET methods, the recurrent layers perform detection in addition to tracking.

In the parametric MSET method, the MUSIC algorithm in addition to the spectrogram requires the knowledge of the number of active sources to estimate their individual DOAs. We obtain this knowledge directly from the reference of the dataset. The frame-wise estimated DOAs of MUSIC MUS_{GT} are in 2D spherical space represented by azimuth and elevation angles. The particle filter with Rao-Blackwellized Monte Carlo data association algorithm, hereafter referred to as PF, processes the frame-wise DOAs to estimate the MSET output MUS^{PF}_{GT}. More details about the PF can be read in [124] and the adaptation to the current task can be read in [VI]. Unlike the parametric method, the input to SELDnet is the phase and magnitude component of the spectrogram of dimension $T \times F \times 2K$ and it produces two outputs: SED of $T \times C$ dimension and MSET output of $T \times 3C$ dimension, together producing the SELDT output as seen in Figure 5.10. Here, the $3C$ represents the 3D Cartesian coordinates of a DOA on a unit sphere.

5.4.2 Evaluation

The SELDnet is evaluated on three static Ambisonic datasets - ANSYN, RESYN and REAL, and two moving source datasets - MANSYN and MREAL (see Section 2.4.1). The SELDnet architecture is unchanged from [V]. Only the sequence length of input frames was tuned for each of the datasets. A 256 frame sequence gave the best results for reverberant datasets, whereas 512 frames gave the best results for anechoic datasets. The PF of the MSET method was tuned on the development split to obtain the best MSET score, before evaluating on the testing split.

The SED and MSET predictions of the SELDnet and the corresponding reference is visualized in Figure 5.11. Each sound class and their corresponding trajectory in azimuth and elevation angles are represented with a unique color. We observe that the SED output is accurate, while the DOA estimations are varying around the reference trajectory with

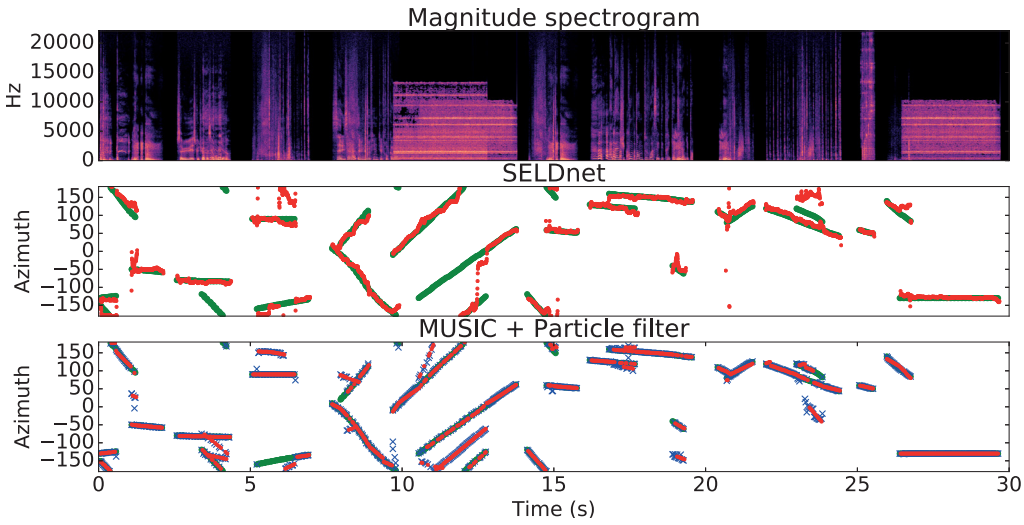


Figure 5.12: Visualization of the MSET results for DNN-based SELDnet and parametric baseline for a recording in MANSYN *O*2 dataset. The top figure shows the input magnitude spectrogram. The center and bottom figures show the output of SELDnet and $\text{MUS}_{\text{GT}}^{\text{PF}}$ tracker in red, and the reference in green. The blue crosses in the bottom figure represent the frame-wise DOA output of MUSIC.

a small deviation. This shows that the SELDnet can successfully perform SELDT of multiple overlapping and moving sources.

The tracking estimates and corresponding reference for SELDnet and the MSET method are visualized in Figure 5.12. The performance of the two methods is visually comparable, with both the methods getting confused in similar situations, for example in the intervals of 4-5 s, 10-12 s, and 23-25 s. The current implementation of SELDnet only estimates one DOA per time-frame for a given sound class. But in a real-life sound scene, there can be multiple instances of the same class overlapping with each other at a given time-frame. The datasets studied in this work also have such same class overlapping frames (SCOF in Table 5.8). In these time-frames, the DOA estimates of the current SELDnet implementation are ambiguous. A similar situation of same class overlapping can be observed in Figure 5.12 in the 10-15 s interval. The SELDnet initially tracks the first source and then starts tracking the second source. The parametric method is seen to track both the overlapping sources. In addition, it produces a false track between the two active sources location. This is a result of MUSIC getting confused between the two identical sources. This confusion can be a result of the known performance degradation of MUSIC in the presence of multiple correlated sources.

The SED and MSET metric scores of SELDnet are tabulated in Table 5.8. To have a direct comparison between the parametric method and SELDnet, the minimum description length (MDL) [35] principle is used to estimate the number of active sources from the input spectrogram. The resulting MUSIC output is represented as MUS_{MDL}. The corresponding MSET output from the PF is represented as MUS_{MDL}^{PF}. This way both the methods use identical spectrogram inputs and produce MSET outputs. Further, in the SELDnet, the recurrent layers were employed to model the long-term temporal structure of the sound events and their trajectories. To study the importance of these recurrent layers for the SELDT task, we train a version of SELDnet without any recurrent layers,

Table 5.8: The evaluation scores of SELDnet and corresponding baselines on different static and moving source datasets. DE: DOA error, FR: Frame recall, F: F-score, SCOF: Same class overlapping frames

		ANSYN			RESYN			REAL			MANSYN			MREAL		
MSET results		*O*1	*O*2	*O*3	*O*1	*O*2	*O*3	*O*1	*O*2	*O*3	*O*1	*O*2	*O*3	*O*1	*O*2	*O*3
MUS_{GT}	DE	1.3	5.0	12.2	21.7	28.9	32.5	15.1	33.9	44.1	0.6	14.8	28.0	16.4	34.1	43.9
MUS^{PF}_{GT}	DE	0.1	1.1	2.3	4.0	5.2	6.1	3.3	8.8	12.0	0.2	4.2	8.0	3.6	8.1	11.9
	FR	97.0	88.5	74.3	83.8	55.6	37.3	93.0	71.0	44.7	98.7	92.3	75.1	91.0	69.9	48.3
Number of active sources estimated directly from input data																
MUS_{MDL}	DE	0.5	14.2	24.0	22.3	31.9	38.5	25.3	36.2	44.1	4.2	17.8	28.5	26.5	35.9	44.9
	FR	93.9	**89.4**	**86.7**	61.7	45.6	52.5	53.6	35.7	**57.5**	63.8	48.1	51.85	53.4	35.2	**58.9**
MUS^{PF}_{MDL}	DE	**0.1**	**4.4**	**7.2**	**6.4**	**10.6**	**12.7**	**9.3**	**10.9**	**13.7**	**3.5**	**6.8**	**8.0**	**13.6**	**11.2**	**13.6**
	FR	96.3	83.5	67.7	52.0	34.1	24.2	52.7	40.1	29.6	64	49.9	39.8	58.7	34.4	27.5
CNN	DE	25.7	25.2	26.9	39.1	35.1	31.4	32.0	34.9	37.1	26.1	25.8	28.2	36.6	39.3	40.2
	FR	80.2	45.6	32.2	69.5	45.8	29.7	45.1	28.4	16.9	83.7	58.1	38.3	44.5	26.2	16.3
SELDnet	DE	3.4	13.8	17.3	9.2	20.2	26.0	26.6	33.7	36.1	6.0	12.3	18.6	36.5	39.6	38.5
	FR	**99.4**	85.6	70.2	**95.8**	**74.9**	**56.4**	**64.9**	**41.5**	24.6	**98.5**	**94.6**	**80.7**	**69.6**	**42.8**	28.9
SED results																
CNN	ER	0.52	0.46	0.51	0.44	0.45	0.54	0.52	0.51	0.51	0.59	0.47	0.48	0.46	0.49	0.52
	F	70.1	66.5	68	57	54.9	42.7	50.1	49.5	48.9	65.6	62.7	60.1	55.4	50.9	48.8
SELDnet	ER	**0.04**	**0.16**	**0.19**	**0.1**	**0.29**	**0.32**	**0.4**	**0.49**	**0.53**	**0.07**	**0.1**	**0.2**	**0.37**	**0.45**	**0.49**
	F	**97.7**	**89**	**85.6**	**92.5**	**79.6**	**76.5**	**60.3**	**53.1**	**51.1**	**95.3**	**93.2**	**87.4**	**64.4**	**56.4**	**52.3**
SCOF (in %)		0.0	4.2	12.1	0.0	4.2	12.1	0.0	7.6	23.0	0.0	3.0	9.1	0.0	7.1	20.9

i.e., we use only the convolutional and fully connected layers, hereafter referred to as CNN. The best CNN architecture had five convolutional layers with 64 filters each.

The overall results from the Table 5.8 are as follows. Across the methods and datasets, the higher the number of overlapping sources, the poorer is the SELDT performance. Usage of temporal filtering of DOAs using PF improves the DOA error across datasets for the parametric method. Usage of MDL instead of the dataset reference for the number of active sources resulted in poor frame recall, especially in reverberant scenes. This indicates that a more robust source detection and counting scheme is required for better performance. The DNN-based SELDnet is observed to perform considerably better than MUS_{MDL} for most datasets. With the usage of PF, the DOA error of MUS^{PF}_{MDL} is observed to be better than SELDnet, while the frame recall remains poorer than SELDnet. In comparison to MUS^{PF}_{GT}, which uses the reference number of active sources, the frame-recall of SELDnet is competitive for all datasets other than the real-life datasets REAL and MREAL, indicating the need for learning from larger datasets and stronger models.

Finally, it was observed that the tracking performance of SELDnet without recurrent layers (CNN) was poor with spurious, high variance DOA estimates. This is also observed in the corresponding DOA errors across datasets in Table 5.8, thus indicating the importance of recurrent layers for the SELDT task.

5.4.3 Contributions and Limitations

In [VI], we presented the first results of a DNN-based SELDT for multiple overlapping and moving sources at varying angular velocities. The MSET performance of the proposed SELDnet method was visually comparable to the parametric method. Numerically, SELDnet achieved a higher frame recall, whereas the parametric method achieved a lower angular error. This is also the first such study on DNN-based MSET with an exhaustive evaluation with datasets comprising of static and moving sources, a varying number of overlapping sources, and anechoic and reverberant scenarios. Additionally, we also showed that recurrent layers are a crucial part for the temporal tracking of sound events and trajectories in the SELDT task. Finally, to support open research and reproducibility, the studied datasets and the methods have been made publicly available[3]. The limitation of SELDnet continues to be the same as discussed in Section 5.3, i.e., poor angular error compared to baseline methods. The localization performances can potentially be improved by learning on larger datasets with more powerful models.

[3]https://github.com/sharathadavanne/seld-net

6 Conclusions and Future Work

In this chapter, we summarize the conclusions from different publications that comprise this thesis. Thereafter, we list the potential future research directions for the SELDT task.

6.1 Conclusions

In this thesis, we investigated the application of DNNs for the SELDT task. In the process, we gained insights on the sub-tasks of SELDT such as DOA estimation, SED and MSET. The SELDT task and the sub-tasks were each studied individually starting from the existing methods. Novel sound representations and corresponding DNN adaptations supporting these representations were proposed that resulted in establishing new state-of-the-art methods for the respective tasks.

In [I], we proposed the first multichannel SED method exploiting the spatial and harmonic features inspired by the human auditory system to recognize overlapping sound events better. A multilayered LSTM network was used as a classifier to map these proposed features to the temporal activity of the corresponding sound event. The results showed that using spatial features from the multichannel audio improves SED in comparison to using single channel audio. The proposed method with multichannel log mel-band energies won the IEEE AASP challenge on SED at DCASE 2016 making it the state-of-the-art at that time.

In [II], we proposed a CRNN network that supports multiple feature classes and can easily scale to features from any number of input channels, while learning the relevant spatial information in the multichannel features. We also showed that the CRNN can learn the relevant information in hand-crafted features of [I] directly from generic low-level features, thus making the feature extraction step independent of the dataset. This method won the IEEE AASP challenge on SED at DCASE 2017 making it the state-of-the-art during that time.

In [III], we showed that complex acoustic scenes with a higher number of overlapping sound events require larger CRNN models. We studied SED performance in an identical polyphonic sound scene with single, binaural and four-channel audio. The results showed that the overlapping sound events are recognized better with a higher number of spatial sampling, i.e., four-channel audio.

In [IV], we presented the first DNN-based method for DOA estimation of multiple temporally overlapping sound events in complete azimuth and elevation space. Unlike the parametric DOA estimators that require the knowledge of the number of active sources, the proposed method learns the number of active sources from the input feature, and estimates their corresponding DOA. In comparison to a parametric baseline approach,

the DOAs estimated by the proposed method had a lower angular error, while the frame recall of DOAs was poor for time-frames with a higher number of overlapping sound events.

In [V], we proposed a SELDT method using CRNN that was that performs SED and MSET jointly. The proposed method is the first DNN-based SELDT method that can localize and track multiple overlapping sound events in a dynamic sound scene with both static and moving sources. Unlike [IV] where the localization was performed as a multiclass multilabel classification, in [V] localization was performed as a multi-output regression task. This enabled the method to estimate continuous and unseen DOA values. The proposed method was shown to be a generic approach to perform SELDT independent of the microphone array architecture. We showed that the method was robust to mismatched reverberant scenarios, hence allowing the reuse of a trained model on comparable reverberant scenarios.

Finally, in [VI], we evaluated the tracking performance of [V] on moving sources at different angular velocities. We showed that the recurrent layers of the SELDT method are crucial for tracking, and that these layers have similar modeling capacity as parametric MSET methods. The MSET performance of the SELDT method was exhaustively compared with parametric MSET methods in multiple acoustic scenes with a different number of overlapping sound events, static and moving sources, and anechoic and reverberant scenarios. It was observed that the proposed method had better frame recall and higher angular error in comparison to the parametric MSET method.

6.2 Future Work

In general, the existing research on the SELDT task, including both the parametric and DNN-based approaches, is limited. Therefore, there are a variety of possible future research directions. We list the important general directions for future research based on our understanding.

Dataset Data-driven approaches to SELDT task require sufficient data. However, annotating real-life recordings for SELDT is a tedious task. One way to overcome this is to build methods that can learn real-life SELDT from large synthesized datasets (accompanied by a smaller real-life evaluation dataset). This requires the synthesized datasets to be as similar to real-life conditions as possible. The work presented in this thesis approached SELDT with a similar idea, by using real-life impulse responses, isolated sound events, and ambient sound to recreate acoustic scenes similar to real-life conditions. However, much remains to be done with respect to developing methods to achieve realistic sound scene synthesis, which in turn could drive future SELDT research. Alternatively, to overcome the annotation problem, weak labeling of the dataset can be explored. The weak labels can be in the form of annotating only the approximate direction of the sound event or annotating just the number of active sound events in a segment without their location.

Method A real-life sound scene can have multiple temporally and spatially overlapping sound events. The proposed method was shown to work successfully on such sound scenes. However, the current architecture is limited in the number of instances of the same class it can detect at a time. Although the proposed method can be extended to localize multiple instances of the same class at a given time, the performance of this extension has not been studied and should be evaluated. Furthermore, sound

events in real-life cannot always be localized to a single coordinate in space. Future methods will have to accommodate diffuse sources such as a car or a train that can extend over a volume or a range of directions. Finally, methods for domain adaptation and learning from weak labels will have to be developed for exploiting synthetic and weakly-labeled datasets for real-life SELDT.

Productization With the surge in smart devices, smart homes and smart cities, SELDT will play a key role to enhance their auditory context-awareness. Task-specific products such as digital assistants enabling sound scene visualization for hearing impaired, or services for monitoring bio-diversity or home and urban environments, can greatly benefit from SELDT. Productization would require methods that are expected to work seamlessly irrespective of the acoustic environment they are used in, and produce results with high confidence. The current methods have not reached this maturity, hence more research in this direction would be required. More about similar productization of machine audition methods for smart homes can be read in [28].

Bibliography

[1] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Workshop on Mobile Computing Systems and Applications*, 1899, pp. 85–90.

[2] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8677–8681.

[3] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM International Conference on Multimedia*, 2014.

[4] S. Adavanne, K. Drossos, E. Cakir, and T. Virtanen, "Stacked convolutional and recurrent neural networks for bird audio detection," in *European Signal Processing Conference*, 2017.

[5] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *International Joint Conference on Neural Networks*, 2017.

[6] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[7] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *IEEE Spoken Language Technology Workshop*, 2016.

[8] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," in *Journal of Robotics and Mechatronics*, vol. 29, no. 1, 2017.

[9] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *International Conference on Robotics and Automation*, 2018.

[10] C. Busso, S. Hernanz, C.-W. Chu, S.-i. Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan, "Smart room: participant and speaker localization and identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[11] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997.

[12] M. Wölfel and J. McDonough, *Distant Speech Recognition*. John Wiley & Sons, 2009.

[13] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," in *IEEE Signal Processing Letters*, vol. 21, 2014.

[14] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *European Signal Processing Conference*, 2011.

[15] "Deafness and hearing loss, key facts," in *World Health Organization*, accessed on 28 Feb 2019. [Online]. Available: http://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss

[16] E. Browning, R. Gibb, P. Glover-Kapfer, and K. E. Jones, "Passive acoustic monitoring in ecology and conservation," in *World Wildlife Fund Conservation Technology Series 1(2)*, accessed on 28 Feb 2019. [Online]. Available: https://www.wwf.org.uk/conservationtechnology/documents/Acoustic monitoring-WWF-guidelines.pdf

[17] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time-frequency audio features," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, 2009.

[18] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack, "Estimating animal population density using passive acoustics," in *Biological Reviews of the Cambridge Philosophical Society*, vol. 88, no. 2, 2012.

[19] B. J. Furnas and R. L. Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," in *Journal of Wildlife Management*, vol. 79, no. 2, 2014.

[20] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," in *Communications of the ACM*, no. 2, 2019, pp. 68–77.

[21] C. Peckens, C. Porter, and T. Rink, "Wireless sensor networks for long-term monitoring of urban noise," *Sensors*, vol. 18, no. 9, p. 3161, 2018.

[22] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of low-cost urban acoustic monitoring devices," *Applied Acoustics*, vol. 117, pp. 207–218, 2017.

[23] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," in *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2008.

[24] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," in *ACM Computing Surveys*, 2016.

[25] C. Grobler, C. Kruger, B. Silva, and G. Hancke, "Sound based localization and identification in industrial environments," in *IEEE Industrial Electronics Society*, 2017.

[26] P. W. Wessels, J. V. Sande, and F. V. der Eerden, "Detection and localization of impulsive sound events for environmental noise assessment," in *The Journal of the Acoustical Society of America 141*, vol. 141, no. 5, 2017.

[27] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, 2015.

[28] S. Krstulović, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 335–371.

[29] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation," *IEEE Transactions on Automation Science and Engineering*, vol. 5, no. 1, pp. 25–31, 2008.

[30] B. Ottersten, M. Viberg, P. Stoica, and A. Nehorai, "Exact and large sample maximum likelihood techniques for parameter estimation and detection in array processing," in *Radar Array Processing. Springer Series in Information Sciences*, 1993.

[31] S. Rickard and F. Dietrich, "DOA estimation of many w-disjoint orthogonal sources from two mixtures using DUET," in *IEEE Workshop on Statistical Signal and Array Processing*, 2000, pp. 311–314.

[32] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *International Symposium on Signal Processing and Its Applications*, 2003.

[33] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.

[34] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, no. 5, pp. 592–605, 2011.

[35] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[36] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[37] S. Chakrabarty and E. A. P. Habets, "Multi-speaker localization using convolutional neural network trained with noise," in *Neural Information Processing Systems*, 2017.

[38] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Springer, 2001.

[39] H. R. Goerlitz, "Weather conditions determine attenuation and speed of sound: Environmental limitations for monitoring and analyzing bat echolocation," *Ecology and Evolution*, vol. 8, no. 10, pp. 5090–5100, 2018.

[40] Y. Bando, T. Mizumoto, K. Itoyama, K. Nakadai, and H. G. Okuno, "Posture estimation of hose-shaped robot using microphone array localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 3446–3451.

[41] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *ACM on Multimedia Conference*, 2016.

[42] A. Kumar and B. Raj, "Deep CNN framework for audio event recognition using weakly labeled web data," in *arXiv:1707.02530v2*, 2017.

[43] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[44] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," in *Workshop on Detection and classification of acoustic scenes and events*, 2017.

[45] S. Adavanne and T. Virtanen, "Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network," in *Workshop on Detection and classification of acoustic scenes and events*, 2017.

[46] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," in *European Signal Processing Conference*, 2018.

[47] A. Kumar, M. Khadkevich, and C. Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[48] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in *European Signal Processing Conference*, 2010.

[49] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *European Signal Processing Conference*, 2016.

[50] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE2017 challenge setup: Tasks, datasets and baseline system," in *Detection and Classification of Acoustic Scenes and Events*, 2017.

[51] E. Benetos, M. Lagrange, and G. Lafay, "Sound event detection in synthetic audio," 2016, accessed on 7 May 2018. [Online]. Available: https://archive.org/details/dcase2016_task2_train_dev

[52] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015.

[53] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. Plumbley, and D. Ellis, Eds. Springer International Publishing, 2018, ch. 6.

[54] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, vol. 6, no. 6, 2016.

[55] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[56] H. W. Kuhn, "The hungarian method for the assignment problem," in *Naval Research Logistics Quarterly*, no. 2, 1955, p. 83 – 97.

[57] D. Gerhard, *Audio Signal Classification: History and Current Techniques.* Citeseer, 2003.

[58] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017.

[59] M. H. Holmes and L. A. Rubenfeld, *Mathematical Modeling of the Hearing Process.* Springer-Verilag, 1980.

[60] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185 – 190, 1937.

[61] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, 2017.

[62] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[63] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," in *Detection and Classification of Acoustic Scenes and Events*, 2017.

[64] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *European Signal Processing Conference*, 2010.

[65] R. Chakraborty and C. Nadeu, "Sound-model-based acoustic source localization using distributed microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[66] L. Lu, F. Ge, Q. Zhao, and Y. Yan, "A SVM-based audio event detection system," in *International Conference on Electrical and Control Engineering*, 2010, pp. 292 – 295.

[67] A. Temko, C. Nadeu, and J.-I. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07," in *Multimodal Technologies for Perception of Humans.* Springer, 2008.

[68] K. Lopatka, J. Kotus, and A. Czyzewsk, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools and Applications Journal*, vol. 75, no. 17, 2016.

[69] E. Çakır, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi-label deep neural networks," in *IEEE International Joint Conference on Neural Networks*, 2015.

[70] P. J. Werbos, "Backpropagation through time: what it does and how to do it," in *Proceedings of the IEEE*, vol. 78, no. 10, 1990.

[71] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[72] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," in *Journal of Machine Learning Research*, 2014.

[73] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, 2015.

[74] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[75] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[76] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[77] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[78] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[79] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017, pp. 374–378.

[80] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[81] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[82] J. Salamon and J. P. Bello, "Feature learning with deep scattering for urban sound analysis," in *European Signal Processing Conference*, 2015.

[83] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[84] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[85] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition," in *Sound and Music Computing Conference*, 2017.

[86] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech and Music Processing*, 2013.

[87] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Workshop on Machine Listening in Multisource Environments, CHiME2011*, Italy, 2011, pp. 36–40.

[88] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[89] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," in *INTERSPEECH*, 2016.

[90] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[91] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Multi-label vs. combined single-label sound event detection with deep neural networks," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 2551–2555.

[92] J. W. Strutt, "On our perception of sound direction," in *Philosophical Magazine*, 1907.

[93] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, 1990.

[94] B. Uzkent, B. D. Barkana, and H. Cevikalp, "Non-speech environmental sound classification using SVM's with a new set of features," in *International Journal of Innovative Computing, Information and Control*, 2012.

[95] B. McFee *et al.*, "librosa v0.4.1," accessed 16.01.2019. [Online]. Available: http://dx.doi.org/10.5281/zenodo.32193

[96] J. O. Smith, *Sinusoidal Peak Interpolation, in Spectral Audio Signal Processing*, accessed 16.01.2019. [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/Sinusoidal_Peak_Interpolation.htm

[97] C. Knapp and C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976.

[98] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *European Signal Processing Conference*, 2016.

[99] "Sound event detection in real life audio." Detection and Classification of Acoustic Scenes and Events 2016, accessed 16.01.2019. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio

[100] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *International Conference on Learning Representations*, 2014.

[101] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[102] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*, 2015.

[103] M. Yiwere and E. J. Rhee, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," in *Journal of Selected Topics in Signal Processing*, vol. 13 (4), no. 22, 2019, pp. 787–799.

[104] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[105] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2016.

[106] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilisitic neural network," in *IEEE Transactions on Industrial Electronics*, vol. 29, no. 1, 2017.

[107] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Deutsche Jahrestagung für Akustik*, 2015.

[108] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," in *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, 2001.

[109] M. S. Brandstein and H. F. Silverman, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997.

[110] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, 1986.

[111] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 37, no. 7, 1989.

[112] D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone-array measurements," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 371–374.

[113] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2002.

[114] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of Multiple Moving Speakers With Multiple Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[115] J. M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216 – 228, 2007.

[116] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 4, pp. 728 – 739, 2008.

[117] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte Carlo data association," in *IEEE Workshop on Statistical Signal Processing*, 2009.

[118] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1409 – 1415, 2012.

[119] J. Traa and P. Smaragdis, "Multiple speaker tracking with the Factorial von Mises-Fisher Filter," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

[120] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 392 – 402, 2014.

[121] J. Nix and V. Hohmann, "Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 995 – 1008, 2007.

[122] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 806 – 815, 2013.

[123] N. Strobel, S. Spors, and R. Rabenstein, *Joint Audio-Video Signal Processing for Object Localization and Tracking.* Springer, 2001, pp. 203 – 225.

[124] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized particle filter for multiple target tracking," *Information Fusion*, vol. 8, no. 1, pp. 2 – 15, 2007.

[125] A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[126] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, 2012.

[127] Y. J. Choe, J. Shin, and N. Spencer, "Probabilistic interpretations of recurrent neural networks," *Probabilistic Graphical Models*, 2017.

Publications

Publication I

Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, Tuomas Virtanen, "Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features," *Detection and Classification of Acoustic Scenes and Events (DCASE).* Budapest, Hungary, pp. 6-10, September 2016.

SOUND EVENT DETECTION IN MULTICHANNEL AUDIO USING SPATIAL AND HARMONIC FEATURES

Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology

ABSTRACT

In this paper, we propose the use of spatial and harmonic features in combination with long short term memory (LSTM) recurrent neural network (RNN) for automatic sound event detection (SED) task. Real life sound recordings typically have many overlapping sound events, making it hard to recognize with just mono channel audio. Human listeners have been successfully recognizing the mixture of overlapping sound events using pitch cues and exploiting the stereo (multichannel) audio signal available at their ears to spatially localize these events. Traditionally SED systems have only been using mono channel audio, motivated by the human listener we propose to extend them to use multichannel audio. The proposed SED system is compared against the state of the art mono channel method on the development subset of TUT sound events detection 2016 database [1]. The usage of spatial and harmonic features are shown to improve the performance of SED.

***Index Terms*—** Sound event detection, multichannel, time difference of arrival, pitch, recurrent neural networks, long short term memory

1. INTRODUCTION

A sound event is a segment of audio that a human listener can consistently label and distinguish in an acoustic environment. The applications of such automatic sound event detection (SED) are numerous; embedded systems with listening capability can become more aware of its environment [2][3]. Industrial and environmental surveillance systems and smart homes can start automatically detecting events of interest [4]. Automatic annotation of multimedia can enable better retrieval for content based query methods [5][6] .

The task of automatic SED is to recognize the sound events in a continuous audio signal. Sound event detection systems built so far can be broadly classified to monophonic and polyphonic. Monophonic systems are trained to recognize the most dominant of the sound events in the audio signal [7]. While polyphonic systems go beyond the most dominant sound event and recognize all the overlapping sound events in a segment [7][8][9][10]. We propose to tackle such polyphonic soundscape which replicates real life scenario in this paper.

Some SED systems have tackled polyphonic detection using mel-frequency cepstral coefficients (MFCC) and hidden Markov models (HMMs) as classifiers with consecutive passes of the Viterbi algorithm [7]. In [11], a non-negative matrix factorization was used as a pre-processing step, and the most prominent event in each of the stream was detected. However, it still had a hard constraint of estimating the number of overlapping events. This was overcome by using coupled NMF in [12]. Dennis et al [8] took an entirely different path from the traditional frame-based features by combining generalized Hough transform (GHT) with local spectral features.

More recently, the state of the art SED systems have used log mel-band energy features in DNN [9], and RNN-LSTM [10] networks trained for multi-label classification. Motivated by the good performance of RNN-LSTM over DNN as shown in [10], we continue to use the multi-label RNN-LSTM network.

The present state of the art polyphonic SED systems have been using a single channel of audio for sound event detection. Polyphonic events can potentially be tackled better if we had multichannel data. Just like humans use their two ears (two channels) to recognize and localize the sound events around them [13], we can also potentially train machines to learn sound events from multichannel of audio. Recently, Xiao et al [14] have successfully used spatial features from multichannel audio for far field automatic speech recognition (ASR) and shown considerable improvements over just using mono channel audio. This further motivates us to use spatial features for SED tasks. In this paper, we propose a spatial feature along with harmonic feature and prove its superiority over mono channel feature even with a small dataset of around 60 minutes.

The remaining of the paper is structured as follows. We describe in Section 2 the features used and the proposed approach. Section 3 presents a short introduction to RNNs and long short-term memory (LSTM) blocks. Section 4 presents the experimental set-up and results on a database of real life recordings. Finally, we present our conclusions in Section 5.

2. SOUND EVENT DETECTION

The sound event detection task involves identifying temporally the locations of sound event and assigning them to one among the known set of labels. Sound events in real life have no fixed pattern. Different contexts, for example, forest, city, and home have a different variety of sound events. They can be of different sparsity based on the context, and can occur in isolation or be completely overlapped with other sound events. While recognizing isolated sounds have been done with an appreciable accuracy [15], detecting the mixture of labels in an overlapped sound event is a challenging task, where still a considerable amount of improvements can be made. Figure 2 shows a snippet of sound event annotation, where three sound events - speech, car, and dog bark happen to occur. At time frame t, two events - speech and car are overlapping. An ideal SED system should be able to handle such overlapping events.

The human auditory system has been successfully exploiting the stereo (multichannel) audio information it receives at its ears to

The research leading to these results has received funding from the European Research Council under the European Unions H2020 Framework Programme through ERC Grant Agreement 637422 EVERYSOUND, and Google Faculty Research Award project “Acoustic Event Detection and Classification Using Deep Recurrent Neural Networks”. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

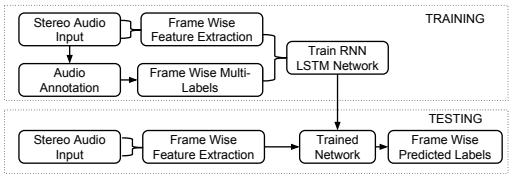


Figure 1: Framework of the training and testing procedure for the proposed system.

isolate, localize and classify the sound events. A similar set up is envisioned and implemented, where the sound event detection system gets a stereo input and suitable spatial features are implemented to localize and classify sound events.

The proposed sound event detection system, shown in Figure 1, works on real life multichannel audio recordings and aims at detecting and classifying isolated and overlapping sound events.

Three sets of features -log mel-band energies, pitch frequency, and its periodicity, and time difference of arrival (TDOA) in subbands, are extracted from the stereo audio. All features are extracted at a hop length of 20 ms to have consistency across features.

2.1. Log mel-band Energy

Log mel-band energies have been used for mono channel sound event detection extensively [9][10][16] and have proven to be good features. In the proposed system we continue to use log mel-band energies, and extract it for both the stereo channels. This is motivated from the idea that human auditory system exploits the interaural intensity difference (IID) for spatial localization of sound source [13]. Neural networks are capable of performing linear operations, which includes the difference. Therefore, when trained on the stereo log mel-band energy data, it will learn to obtain information similar to IID.

Each channel of the audio is divided into 40 ms frames with 50% overlap using hamming window. Log mel-band energies are then extracted for each of the frames (mel in Table 1). We use 40 mel-bands spread across the entire spectrum.

2.2. Harmonic features

The pitch is an important perceptual feature of sound. Human listeners have evolved to identify different sounds using the pitch cues, and can make efficient use of pitch to acoustically separate each of the mixture in an overlapping sound event [17]. Uzkent et al [18] have shown improvement in accuracy of non speech environmental sound detection used pitch range along with MFCC's. Here we propose using the absolute pitch and its periodicity as the features ($pitch$ in Table 1).

The librosa implementation of pitch tracking [19] on thresholded parabolically-interpolated STFT [20] was used to estimate the pitch and periodicity.

Since we are handling multi-label classification it is intuitive to identify as many dominant fundamental frequencies as possible and use them to identify the sound events. The periodicity feature gives the confidence measure for the extracted pitch value and helps the classifier to make better decisions based on pitch.

The overlapping sound events in the training data (Section 4.1) did not have more than three events overlapping at a time, hence we have limited ourselves to using the top three dominant pitch values per frame. So, for each of the channels, top three pitch values, and its respective periodicity values are extracted at every frame in 100-4000 Hz frequency range ($pitch3$ in Table 1).

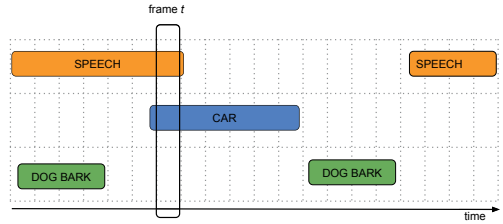


Figure 2: Sound events in a real life scenario can occur in isolation or overlapped. We see that at frame *t*, speech and car events are overlapping.

2.3. Time difference of arrival (TDOA) features

Overlapping sound events have forever troubled classification systems. This is mainly because the feature vector for the overlapped frame is a combination of different sound events. But, human listeners have been able to successfully identify each of the overlapping sound events by isolating and localizing the source spatially. This has been possible due to the interaural time delay (ITD) [13]

Each sound event has its own frequency band, some occur in low frequencies, some in high, and some occur all across the frequency band. If we can divide the frequency spectrum into different bands, and identify the spatial location of the sound source in each of these bands, then this is an extra dimension of the feature, which the classifier can learn to estimate the number of possible sources in each frame, and their orientation in the space. We implement this by dividing the spectral frame into five mel-bands and calculating the time difference of arrival (TDOA) at each of these bands.

For example, if a non-overlapping isolated sound event is spread across the entire frequency range, and we are calculating the TDOA in five mel-bands. We should have the same TDOA values for each of the bands. However, if we have two overlapping sounds S_1 and S_2, where S_1 is spread in the first two bands and S_2 is spread in the last two bands. The feature vector will have different TDOA values for each of the sounds, which the classifier can learn to isolate and identify them as separate sound events.

The TDOA can be estimated using the generalized cross-correlation with phase-based weighting (GCC-PHAT) [21]. Here, we extract the correlation for each mel-band separately:

$$R_b(\Delta_{12}, t) = \sum_{k=0}^{N-1} H_b(k) \frac{X_1(k,t) \cdot X_2^*(k,t)}{|X_1(k,t)||X_2(k,t)|} e^{i2\pi k \Delta_{12}/N}, \quad (1)$$

where N is the number of frequency bands, $X(k,t)$ is the FFT coefficient of the kth frequency band at time frame t and the subscript specifies the channel number, $H_b(k)$ is the magnitude response of the bth mel-band of total of B bands and Δ_{12} is the sample delay value between channels. The TDOA is extracted as the location of correlation peak magnitude for each mel-band and time frame.

$$\tau(b,t) = \underset{\Delta_{12}}{\operatorname{argmax}} \{R_b(\Delta_{12}, t)\} \quad (2)$$

The maximum and minimum TDOA values are truncated between values $-2\tau_{\text{max}}, 2\tau_{\text{max}}$, where τ_{max} is the maximum sample delay between a sound wave traveling between microphones.

Feature Name	Length	Description
mel	40	Log mel-band energy extracted on a single channel of audio
$pitch$	2	Most dominant pitch value and periodicity extracted on a single channel
$pitch3$	6	Top three dominant pitch and periodicity values extracted on a single channel
$tdoa$	5	Median of multi-window TDOA's extracted from stereo audio
$tdoa3$	15	Concatenated multi-window TDOA's extracted from stereo audio

Table 1: Definitions of acoustic features proposed for sound event detection.

The sound events in the training set were seen to be varying from 50 ms to a few seconds. In order to accommodate such variable length sound events, TDOA was calculated in three different window lengths — 120, 240 and 480 ms, with a constant hop length of 20 ms. The TDOA values of these three windows were concatenated for each mel-band to form one set of TDOA features. So, TDOA values extracted in five mel-band, and for three window lengths, on concatenation gives 15 TDOA values per frame ($tdoa3$ in Table 1).

TDOA values in small windows are generally very noisy and unreliable. To overcome this, the median of the TDOA values from the above three different window lengths for each sub-band of the frame was used as the second set of TDOA features ($tdoa$ in Table 1). Post filtering across window lengths, the TDOA values in each mel-band were also median filtered temporally using a kernel of length three to remove outliers.

3. MULTI-LABEL RECURRENT NEURAL NETWORK BASED SOUND EVENT DETECTION

Deep neural networks have shown to perform well on complex pattern recognition tasks, such as speech recognition [22], image recognition [23] and machine translation [24]. A deep neural network typically computes a map from an input to an output space through several subsequent matrix multiplications and non-linear activation functions. The parameters of the model, i.e. its weights and biases, are iteratively adjusted using a form of optimization such as gradient descent.

When the network is a directed acyclic graph, i.e. information is only propagated forward, it is known as a feedforward neural network (FNN). When there are feedback connections the model is called a recurrent neural network (RNN). An RNN can incorporate information from previous timesteps in its hidden layers, thus providing context information for tasks based on sequential data, such as temporal context in audio tasks. Complex RNN architectures — such as long short-term memory (LSTM) [25] — have been proposed in recent years in order to attenuate the vanishing gradient problem [26]. LSTM is currently the most widely used form of RNN, and the one used in this work as well.

In SED, RNNs can be used to predict probabilities for each class to be active in a given frame at timestep t. The input to the network is a sequence of feature vectors $\mathbf{x}(\mathrm{t})$; the network computes hidden activations for each hidden layer, and at the output layer a vector of predictions for each class $\mathbf{y}(\mathrm{t})$. A sigmoid activation function is used at the output layer in order to allow several classes to be predicted as active simultaneously. By thresholding the predictions at the output layer it is possible to obtain a binary activity matrix.

3.1. Neural network configurations

For each recording, we obtain a sequence of feature vectors, which is normalized to zero mean and unit variance, and the scaling parameters are saved for normalizing the test feature vectors. The sequences are further split into non-overlapping sequences of length 25 frames. Each of these frames has a target binary vector, indicating which classes are present in the feature vector.

We use a multi-label RNN-LSTM with two hidden layers each having 32 LSTM units. The number of units in the input layer depends on the length of the feature being used. The output layer has one neuron for each class. The network is trained by back propagation through time (BPTT) [27] using binary cross-entropy as loss function, Adam optimizer [28] and block mixing [10] data augmentation. Early stopping is used to reduce over-fitting, the training is halted if the segment based error rate (ER) (see Section 4.2) on the validation set does not decrease for 100 epochs.

At test time we use scaling parameters estimated on training data to scale the feature vectors and present them in non-overlapping sequences of 25 frames, and threshold the outputs with a fixed threshold of 0.5, i.e., we mark an event is active if the posterior in the output layer of network is greater than 0.5 and otherwise inactive.

4. EVALUATION AND RESULTS

4.1. Dataset

We evaluate the proposed SED system on the development subset of TUT sound events detection 2016 database [1]. This database has stereo recordings which were collected using binaural Soundman OKM II Klassik/studio A3 electret in-ear microphones and Roland Edirol R09 wave recorder using 44.1 kHz sampling rate and 24-bit resolution. It contains two contexts - home and residential area. Home context has 10 recordings with 11 sound event classes and the residential area context has 12 recordings with 7 classes. The length of these recordings is between 3-5 minutes.

In the development subset provided, each of the context data is already partitioned into four folds of training and test data. The test data was collected such that each recording is used exactly once as the test, and the classes in it are always a subset of the classes in the training data. Also, 20% of the training data recordings in each fold were selected randomly to be used as validation data. The same validation data was used across all our evaluations.

4.2. Metrics

We perform the evaluation of our system in a similar fashion as [1] which uses the established metrics for sound event detection defined in [30]. The error rate (ER) and F-scores are calculated on one second long segments. The results from all the folds are combined to produce a single evaluation. This is done to avoid biases caused due to data imbalance between folds as discussed in [31].

4.3. Results

The baseline system for the dataset [1] uses 20 static (excluding the 0th coefficient), 20 delta and 20 acceleration MFCC coefficients

	Feature combination	Home		Residential area		Average	
		ER	F (%)	ER	F (%)	ER	F (%)
Baseline system using GMM classifier in DCASE 2016 [1][29]	$mfcc; delta; acc$	0.96	15.9	0.86	31.5	0.91	23.7
Mono channel feature With RNN-LSTM network	mel_1	**0.94**	**27.4**	0.88	38.3	0.91	32.9
Hybrid (mono and stereo) features with RNN-LSTM network	$mel_1; pitch_1$	0.97	25.4	0.85	43.4	0.91	34.4
	$mel_1; pitch3_1$	0.96	27.6	0.88	43.9	0.92	35.7
	$mel_1; tdoa$	1.02	19.4	0.89	40.2	0.96	29.8
	$mel_1; tdoa3$	0.98	25.9	0.87	40.5	0.92	33.2
Stereo features with RNN-LSTM network	mel_2	1.03	25.4	0.84	45.9	0.93	35.6
	$mel_2; pitch_2$	1.03	24.9	0.93	40.9	0.98	32.9
	$mel_2; pitch3_2$	0.97	26.6	0.88	41.7	0.92	34.2
	$mel_2; tdoa$	1.01	24.4	**0.82**	**46.4**	**0.91**	**35.4**
	$mel_2; tdoa3$	0.96	24.9	0.86	38.5	0.91	31.7
	$mel_2; tdoa3; pitch_2$	0.97	25.7	0.85	43.1	0.91	34.4
	$mel_2; tdoa3; pitch3_2$	0.99	26.5	0.91	35.2	0.95	30.9
	$mel_2; tdoa; pitch_2$	0.98	24.7	0.87	43.8	0.92	34.2
	$mel_2; tdoa; pitch3_2$	0.94	26.3	0.89	40.5	0.91	33.4

Table 2: Segment based error rate (ER) and F-score achieved for different feature combinations in home and residential area contexts for the development set. The features listed in Table 1 are used in different combinations with the proposed RNN-LSTM network. The subscripts '$_1$' and '$_2$' in the feature combinations column represent how many channels the features were extracted on. For example, feature combination $mel_2; tdoa; pitch_2$ means that the final feature vector has log mel-band energies, most dominant pitch and periodicity values extracted on both the stereo channels, and the time difference of arrival (TDOA) calculated between the stereo channels. The highlighted ER and F-score pair for each context is the best ER score achieved.

extracted on mono audio with 40 ms frames and 20 ms hop length. A Gaussian mixture model (GMM) consisting of 16 Gaussians is then trained for each of the positive and negative values of the class. This baseline system gives a context average ER of 0.91 and F-score of 23.%. An ideal system should have an ER of 0 and an F-score of 100%.

In Table 2 we compare the segment based ER and F-score for different combinations of proposed spatial and harmonic features. In all these evaluations, only the size of the input layer changes based on the feature set, with the rest of the configurations in the RNN-LSTM network remaining unchanged.

Mono channel audio was created by averaging the stereo channels in order to compare the performance of the proposed spatial and harmonic features for multichannel audio. One of the present state of the art SED system for mono channel is proposed in [10]. An RNN-LSTM network is trained in a similar fashion with log mel-band energy feature (Section 2.1) and evaluated. Across contexts, the F-score was seen to be better than the GMM baseline system with comparable ER. Here onwards we use this mono-channel log mel-band feature and RNN-LSTM network configuration result as a baseline for comparisons.

A set of hybrid combinations were tried as shown in Table 2. All combinations other than $mel_1; tdoa$ performed better than the baseline across contexts in F-score.

Finally, the full spectrum of proposed spatial and harmonic features were evaluated in different combinations with RNN-LSTM network. With a couple of exceptions - $mel_2; pitch_2$ and $mel_2; tdoa3; pitch3_2$, all the combinations of features performed equal to or better than the baseline in average F-scores, with marginally similar average ER as baseline. Given the dataset size of around 60 minutes, it is difficult to conclusively say that the binaural features are far superior to monaural features; but they surely look promising.

Binaural features - mel_2 and $mel_2; tdoa; pitch_2$ in Table 3 were submitted to the DCASE 2016 challenge [29], where they were evaluated as the top performing systems. Monaural feature mel_1 was submitted unofficially to compare the performance with binaural features. The hyper-parameters of the network were tuned before the submission, and hence the development set results in Table 3 are different from Table 2. Three hidden layers with 16 LSTM units each were used for mel_2, while mel_1 and $mel_2; tdoa; pitch_2$ were trained with two layers each having 16 LSTM units.

Feature combination	Evaluation dataset		Development dataset	
	ER	F (%)	ER	F (%)
mel_1	0.79	46.6	0.90	35.3
mel_2	0.80	47.8	0.88	34.7
$mel_2; tdoa; pitch_2$	0.88	37.9	0.87	34.8

Table 3: Comparison of segment based error rate (ER) and F-score for development and evaluation dataset. The evaluation dataset scores are the result of DCASE 2016 challenge [29].

5. CONCLUSION

In this paper, we proposed to use spatial and harmonic features for multi-label sound event detection along with RNN-LSTM networks. The evaluation was done on a limited dataset size of 60 mins, which included four cross validation data for two contexts — home and residential area. The proposed multi-channel features were seen to be performing substantially better than the baseline system using mono-channel features.

Future work will concentrate on finding novel data augmentation techniques. Augmenting spatial features is an unexplored space, and will be a challenge worth looking into. Concerning the model, further studies can be done on different configurations of RNN like extending them to bidirectional RNN's and coupling with convolutional neural networks.

6. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.

[2] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with timefrequency audio features," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, 2009, p. 1142.

[3] S. Chu, S. Narayanan, C. C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *IEEE Int. Conf. Multimedia and Expo (ICME)*, 2006, p. 885.

[4] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[5] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 4, no. 2, 2008, p. 11.

[6] D. Zhang and D. Ellis, "Detecting sound events in basketball video archive," in *Dept. Electronic Eng., Columbia Univ., New York*, 2001.

[7] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," in *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, 2013, p. 1.

[8] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform," in *Pattern Recognition Letters*, vol. 34, no. 9, 2013, p. 1085.

[9] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi-label deep neural networks," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2015.

[10] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[11] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in *Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada*, 2013,, p. 8677.

[12] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013,, p. 1.

[13] J. W. Strutt, "On our perception of sound direction," in *Philosophical Magazine*, vol. 13, 1907,, p. 214.

[14] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and DongYu, "Deep beamforming networks for multi-channel speech recognition," in *ICASSP*, 2016,.

[15] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *European Signal Processing Conference (EUSIPCO 2014)*, 2014,.

[16] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[17] A. S. Bregman, "Auditory scene analysis: The perceptual organization of sound," in *MIT Press, Cambridge, MA*, 1990.

[18] B. Uzkent, B. D. Barkana, and H. Cevikalp, "Non-speech environmental sound classification using svms with a new set of features," in *International Journal of Innovative Computing, Information and Control*, 2012, p. 3511.

[19] B. McFee, M. McVicar, C. Raffel, D. Liang, O. Nieto, E. Battenberg, J. Moore, D. Ellis, R. YAMAMOTO, R. Bittner, D. Repetto, P. Viktorin, J. F. Santos, and A. Holovaty, "librosa: 0.4.1," Oct. 2015. [Online]. Available: http://dx.doi.org/10.5281/zenodo.32193

[20] J. O. Smith, *Sinusoidal Peak Interpolation, in Spectral Audio Signal Processing*, accessed 23.06.2016, online book, 2011 edition. [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/Sinusoidal_Peak_Interpolation.htm

[21] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[22] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6645–6649.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[27] P. J. Werbos, "Backpropagation through time: what it does and how to do it," in *Proceedings of the IEEE,*, vol. 78 no. 10, 1990, p. 15501560.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv:1412.6980 [cs.LG]*, December, 2014.

[29] "Detection and classification of acoustic scenes and events," 2016. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio

[30] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, vol. 6(6):162, 2016.

[31] G. Forman and M. Scholz, "Apples-to-apples in cross validation studies: Pitfalls in classifier performance measurement," in *SIGKDD Explor. Newsl.*, vol. 12, no. 1, Nov. 2010, p. 49.

Publication II

Sharath Adavanne, Pasi Pertilä, Tuomas Virtanen, "Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, USA, pp. 771-775, March 2017.

SOUND EVENT DETECTION USING SPATIAL FEATURES AND CONVOLUTIONAL RECURRENT NEURAL NETWORK

Sharath Adavanne, Pasi Pertilä, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology

ABSTRACT

This paper proposes to use low-level spatial features extracted from multichannel audio for sound event detection. We extend the convolutional recurrent neural network to handle more than one type of these multichannel features by learning from each of them separately in the initial stages. We show that instead of concatenating the features of each channel into a single feature vector the network learns sound events in multichannel audio better when they are presented as separate layers of a volume. Using the proposed spatial features over monaural features on the same network gives an absolute F-score improvement of 6.1% on the publicly available TUT-SED 2016 dataset and 2.7% on the TUT-SED 2009 dataset that is fifteen times larger.

Index Terms— Sound event detection, multichannel audio, spatial features, convolutional recurrent neural network

1. INTRODUCTION

Sound event detection (SED) task involves recognizing the onset and offset of a sound event in an acoustic scene and further labeling the sound event. The world we live in offers a rich variety of sound events. For example, recognizing environmental sounds [1][2] will give an idea about the local biodiversity. Detecting sound events such as glass breaking and alarm detection can be used for surveillance [3][4]. Furthermore, the detected sound events can be used as a mid-level representation to help retrieval of content based query [5].

Traditionally SED systems have been using monaural audio. Temko et al. [6] proposed to use multichannel audio, and combined classification likelihoods across channels. While the multichannel audio was used, the actual potential of multichannel features was not exploited. Features like time difference of arrival (TDOA) and mel-band energies from the multichannel audio can potentially help the system differentiate the overlapping sound events. Similar multichannel features have been proposed in automatic speech recognition (ASR) [7] and source separation [8]. Just like humans have evolved to exploit the spatial data available at their ears (multichannel) to identify both isolated and polyphonic sound events [9], we can potentially train the SED systems to learn similar spatial information with multichannel data. Recently, such spatial features motivated by the binaural hearing of humans were proposed and shown to be promising for SED task in [10]. Although the features showed improvement over monoaural features, the dataset was too small (around one hour) to conclusively prove the superiority of binaural spatial features (referred as binaural features in future).

In this paper, we propose to use low-level features and compare it with using high-level features. For example, we compare using generalized cross-correlation with phase based weighting ($GCC\text{-}PHAT$) instead of the high-level TDOA feature which is extracted from $GCC\text{-}PHAT$, and show that the network learns powerful representation from just the low-level features. We show that arranging features from each channel as different layers of a multi-layered input volume enables the network to learn the sound events in multichannel audio better than a simple concatenation of the features. We propose to extend the convolutional recurrent neural network (CRNN) to handle more than one feature type and use a bi-directional LSTM. Finally, we evaluate the improvement of using binaural over monaural features on the 19 hours large TUT-SED 2016 dataset.

We present the binaural features used for SED in Section 2, the extended CRNN architecture in Section 3, the experimental set-up and results on two different real-life datasets in Section 4 and our conclusions in Section 5.

2. BINAURAL FEATURES FOR POLYPHONIC SED

Polyphonic SED is the task of recognizing overlapped sound events along with the isolated sound events. The proposed polyphonic SED system has two parts, feature extraction, and a neural network. The neural network described in Section 3 outputs a vector for every sound event class, where each entry in the vector indicates if the sound event was active or not. The feature extraction part extracts the following binaural features at a constant hop length of 20 ms.

2.1. Binaural mel-band energies

Sound sources which have different spatial locations have different intensities in the binaural channels. Furthermore, most overlapping sound events have different frequency spread in

The research leading to these results has received funding from the European Research Council under the European Unions H2020 Framework Programme through ERC Grant Agreement 637422 EVERYSOUND. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

the spectrum. The combination of this intensity difference in different bands of frequencies can be exploited to differentiate overlapping sound events. This idea is motivated from the interaural intensity difference (IID) used by humans [9].

Log mel-band energies (referred as mel in future) extracted from both of the binaural channels using 40 mel-bands in 40 ms Hamming window are used as the features. A neural network which is capable of performing linear operations, which includes the difference, can learn to obtain the IID information from these channel-wise energies. By using the channel-wise energies instead of the multichannel energy difference directly, we allow the network to learn other potentially more informative features.

2.2. Time difference of arrival vs cross-correlation

Based on how the sound sources are spatially located with respect to the binaural microphones, they might have different $TDOA$ values. Furthermore, sound events which are overlapping do not always have the same frequency spread in the spectrum. The combination of this $TDOA$ difference in different frequency bands can be exploited by a network to differentiate overlapping sound events. We implemented it by dividing the spectral frame into five mel-bands and calculating the $TDOA$ values in each of the bands. The $TDOA$ is estimated using the $GCC\text{-}PHAT$ [11]. The $GCC\text{-}PHAT$ for each mel-band b is extracted separately:

$$R_b(\Delta_{12}, t) = \sum_{k=0}^{N-1} H_b(k) \frac{X_1(k,t) \cdot X_2^*(k,t)}{|X_1(k,t)||X_2(k,t)|} e^{\frac{i2\pi k \Delta_{12}}{N}}, \quad (1)$$

where, X_1 and X_2 are the FFT coefficients of the two binaural channels. $X_1(k,t)$ specifies the coefficient at time frame t and kth frequency bin, of the total N bins. $H_b(k)$ is the magnitude response of the bth band in B mel-bands and $\Delta_{12} \in [-\tau_{\text{max}}, \tau_{\text{max}}]$, where $\tau_{\text{max}} = 30$ is the maximum sample delay for a sound wave to travel between binaural microphones. Finally, the peak magnitude for each mel-band and time frame is picked in the $GCC\text{-}PHAT$ by $\tau(b,t) = \underset{\Delta_{12}}{\text{argmax}} \{|R_b(\Delta_{12}, t)|\}$.

$TDOA$'s for each band are extracted using multi-resolution windows of 120 ms, 240 ms, and 480 ms to accommodate sound events of variable length. Five $TDOA$ values picked from five bands, for each of the three resolutions, results in 15 $TDOA$ values per time frame.

Neural networks have the potential to learn powerful representations from the raw data. We investigate this by using low-level $GCC\text{-}PHAT$ and comparing it with high-level $TDOA$ feature (which are picked from the $GCC\text{-}PHAT$). $GCC\text{-}PHAT$'s are extracted using Eq. 1 with B set to one. To have a factorizable feature length for max pooling, 60 $GCC\text{-}PHAT$ values are picked in -29 to +30 lag for each of the three multi-resolution (same as $TDOA$), amounting to 180 $GCC\text{-}PHAT$ values per time frame. By using $GCC\text{-}PHAT$ instead of $TDOA$, we take the data-oriented approach and get rid of empirical limitations and let the network learn the representation best suited for the problem.

2.3. Dominant frequencies vs auto-correlation

In [10], it was shown that the three most dominant frequencies and their magnitudes (referred as $dom\text{-}freq$ in future) helped in the SED task. This was motivated by the idea that overlapping sound events do not always have the same dominant frequencies, and the network can learn to differentiate these overlapped events using the dominant frequencies. The $dom\text{-}freq$ values were picked from thresholded parabolically-interpolated STFT [12] in the 100 to 4000 Hz range from each of the binaural channels in frames of 40 ms. We continue to use this feature in this paper.

The pitch is a perceptual feature which human listeners have been using to recognize overlapping sound events [13]. One of the prominent way to estimate pitch values are from the auto-correlation (ACR). In the presented work, ACR is calculated on the binaural channels by time domain auto-correlation in 40 ms windows and choosing 400 correlation values in the range of 107.5 Hz to 4410 Hz. This was selected to be close to the $dom\text{-}freq$ extraction range and the number of correlation values easily factorizable during max pooling.

3. CONVOLUTIONAL RECURRENT NEURAL NETWORK

The best results to date in polyphonic SED was reported in [14], where an architecture exploiting the combined modeling capacities of a convolutional neural network (CNN), recurrent neural network (RNN) and fully connected (FC) layer termed as the convolutional recurrent neural network (CRNN) was proposed. We use this CRNN network and extend it for multichannel audio features.

Features from each channel of the multichannel features are layered one over the other to form a volume. More concretely, M frames of a feature, each of length L, from two channels are layered into a $M \times L \times 2$ volume. On slicing such a volume along a particular time frame, we get all the multichannel features corresponding to that time frame. The two-dimensional CNN's by design are built to learn on such volumes, i.e., it initially learns channel-wise filter weights, and further builds an activation map that is obtained as a combination of these channel-wise filter weights, which serves as the inter-channel information. This way we enable the CNN layers in the initial stages of the CRNN network to learn inter-channel information from multichannel features. We report the improvement in performance of using such a volume input over simple multichannel feature concatenation ($M \times 2L$) in Section 4.4.

Separate volumes of each of the multichannel features are created. T time frames of 40 mel features from the two binaural channels are layered into one volume of size $T \times 40 \times 2$. When using $dom\text{-}freq$, dominant frequencies and their magnitudes are treated as different features, and since their feature lengths are the same (3) we layer them in $T \times 3 \times 4$. For ACR we layer the 400 correlation values of each channel into a $T \times 400 \times 2$ volume. Similarly, the three multi-resolution

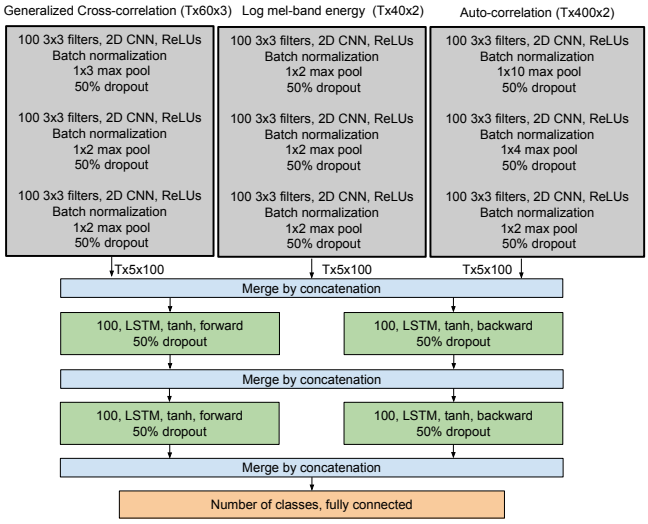


Fig. 1. Convolutional bi-directional recurrent neural network (CBRNN) architecture for multichannel audio features

$TDOA$ features are layered to $T \times 5 \times 3$ and the 60 values of $GCC\text{-}PHAT$ are layered to $T \times 60 \times 3$.

Separate CNN's are used to learn local shift-invariant features in each of these volumes as shown in Figure 1. Since the dimensions of mel, $GCC\text{-}PHAT$, and ACR are high, we use three CNN layers followed by max pooling to reduce the final feature map dimension to $T \times 5 \times 100$. When using $TDOA$ and $dom\text{-}freq$ features, a single 100-filter CNN layer is used without max pooling. To keep the time information intact for final sound event onset and offset detection, we do not apply max pooling in time (T) axis. Post CNNs, the feature maps are merged using concatenation and fed to two consecutive bi-directional long short term memory (LSTM). The output layer is a fully-connected time distributed layer which has as many units as the number of classes in the dataset. A sigmoid activation function is used at the output layer to allow several classes to be predicted as active simultaneously. We refer to this as the CBRNN system in future.

Batch normalization [15] is used in all the CNN layers. A 50% dropout [16] is utilized in all CNNs and LSTMs to avoid over-fitting of the network. The combined architecture was trained by backpropagation through time [17] using Adam optimizer [18] and binary cross-entropy objective. Early stopping was used to reduce overfitting if the F-score (Section 4.2) did not change for 50 epochs. A sequence length of 100 frames (2 seconds) and a batch size of 32 was chosen after calibrating. At test time the sigmoid layer outputs are thresholded with a fixed value of 0.5.

4. EVALUATION AND RESULTS

4.1. Datasets

The proposed SED system is evaluated on two real-life datasets -TUT Sound Events 2009 (TUT-SED 2009) [19] and TUT Sound Events 2016 Development set (TUT-SED 2016) [20]. Both datasets have been recorded using in-ear microphones. TUT-SED 2009 has been used for SED in monaural context [14], but no previous work has reported using the binaural recordings on this dataset. TUT-SED 2016 was published as part of the DCASE 2016 challenge [21], to allow public benchmarking. TUT-SED 2009 is fifteen times larger than TUT-SED 2016, by showing considerable improvement on TUT-SED 2009 we can conclusively say the proposed system is learning and exploiting spatial information.

All the work proposed in this paper is done in a context-independent manner, i.e., we train a single system to learn sound event classes across contexts.

The first dataset - TUT-SED 2009 consists of 103 binaural recordings from 10 different contexts (listed in Table 2). Each context consists of 8 to 14 recordings which vary from 10 to 30 minutes, amounting to an overall length of 1133 minutes. The recordings have been manually annotated, and the annotated events have been grouped into 61 event classes [19]. Each context has 9-16 event classes, while some events occur in multiple contexts, some are context specific. The dataset defines five-folds for training, validation, and testing.

The second dataset - TUT-SED 2016 consists of 22 binaural recordings for two contexts - home and residential area, amounting to 78 minutes. The home context has ten recordings with 11 sound event classes, and the residential area has 12 recordings with seven sound event classes [20]. The dataset defines four-folds for training and testing. We use 20% of the training data for validation, and the same validation is used for all our evaluations.

4.2. Metrics

The SED system output is evaluated with the reference in fixed length intervals, also called as segment-based evaluation [22]. For each segment k, the following are calculated (i) true positive ($TP(k)$): total number of events active in both reference and system output segment. (ii) False positive ($FP(k)$): total number of events active in system output segment but not in reference. (iii) False negative ($FN(k)$): total number of events active in reference segment but not in system output. The first metric, F-score is then calculated as,

$$F = \frac{2 \cdot \sum_{k=1}^{K} TP(k)}{2 \cdot \sum_{k=1}^{K} TP(k) + \sum_{k=1}^{K} FP(k) + \sum_{k=1}^{K} FN(k)} \quad (2)$$

The second metric, error rate (ER) evaluates the system output based on the number of insertions (I), deletions (D) and substitutions (S).

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \quad (3)$$

Where $N(k)$ is the number of sound events marked as active in the reference segment k, and

$$S(k) = \min(FN(k), FP(k)) \quad (4)$$

$$D(k) = \max(0, FN(k) - FP(k)) \quad (5)$$

$$I(k) = \max(0, FP(k) - FN(k)) \quad (6)$$

We use a segment length of one second for ER and F-score estimation. The evaluation metrics are calculated for each context separately and averaged result is presented.

4.3. Baseline

The proposed CBRNN architecture with binaural features is compared with the state of the art monaural SED system introduced in [14]. The system used 40 monaural log mel-band energies (mel-$monaural$) as features. The network had three CNN's each of 96 filters, followed by max pooling in frequency axis reducing the dimension to one. The feature map from CNN was then fed to three LSTMs with 256 units each. The output was a fully-connected layer with units equal to the number of classes in the dataset.

4.4. Results

Table 1 shows the metrics for multi-layered input of the binaural log mel-band energy features (mel) and concatenating it (mel-$concat$) for TUT-SED 2009 dataset. Using a multi-layered input is seen to perform relatively better than a simple concatenation. Similar improvement was observed using multi-layered input of $TDOA$, dom-$freq$, GCC-$PHAT$ and ACR (not tabulated).

From Table 1 we see that using binaural features improves both the ER and F-scores over monaural features (mel-$monaural$) across datasets. While the dom-$freq$ and mel feature combination gave the best performance in TUT-SED 2009, $TDOA$ and mel performed the best for TUT-SED 2016. In numbers, using binaural over monaural features on the same network gives an absolute F-score improvement of 2.7% for TUT-SED 2009 and 6.1% for TUT-SED 2016. By showing this improvement on a larger dataset like TUT-SED 2009, we can more confidently say that the network is truly learning the binaural information.

From the metrics in Table 1 and 2 we see that the performance of using GCC-$PHAT$ instead of $TDOA$ or ACR instead of dom-$freq$, is comparable. This is a significant result, showing that the network can learn equivalent information of powerful high-level features from just the low-level features. Thereby making the features dataset independent and relieving the tuning of parameters like the number of dom-$freq$ and $TDOA$ values.

Most of the sound event classes were seen to be recognized better with the binaural features. Since we cannot present all the 79 classes of the two datasets in this paper, we show the context based F-scores for TUT-SED 2009 dataset in Table 2. A general observation is that the dom-$freq$ / ACR and mel are useful for indoor and sound intense environment (bus, hallway, office, and basketball), while $TDOA$ / GCC-$PHAT$ and mel are seen to help in outdoor contexts (beach and street). This also explains why dom-$freq$ and mel gave better results for TUT-SED 2009. While TUT-SED 2016 had one each of indoor and outdoor contexts, TUT-SED 2009 had more indoor contexts than outdoor.

Feature combination	TUT-SED 2009		TUT-SED 2016	
	ER	F	ER	F
CRNN baseline [14]	0.49	68.8	**0.93**	31.3
mel-$monaural$	0.49	68.0	1.03	29.7
mel-$concat$	0.44	70.3		
mel	**0.43**	71.1	0.99	32.3
mel + $TDOA$	0.45	70.9	0.95	**35.8**
mel + GCC-$PHAT$	0.44	71.1	0.95	34.6
mel + dom-$freq$	**0.43**	**71.7**	0.98	32.8
mel + ACR	0.44	71.2	0.98	33.8
mel + $TDOA$ + dom-$freq$	0.44	71.0	1.01	33.3
mel + GCC-$PHAT$ + ACR	0.45	70.9	0.99	33.6

Table 1. Error rate (ER) and F-score achieved using binaural features and CBRNN on TUT-SED 2009 and 2016 datasets.

The proposed CBRNN architecture using the same mel-$monaural$ feature used in CRNN-baseline achieved an F-score of 68.0% for TUT-SED 2009 and 29.7% for TUT-SED 2016 (Table 1). The difference in the scores with respect to CRNN-baseline can be associated with using a higher dimensional input to LSTM's in the proposed CBRNN.

5. CONCLUSION

In this paper, we extended convolutional recurrent neural networks to handle multiple feature classes and process feature-maps using bi-directional LSTM's. A multi-layered input of multichannel features which enables the network to learn sound events in a multichannel audio better was proposed. Low-level features were used in place of high-level features, and the network was shown to learn high-level equivalent information from simple low-level features. The performance of the system was evaluated on two datasets - a larger dataset for proving that the binaural features truly help in improving the sound event detection, and a public dataset, to allow other researchers to benchmark. The proposed network using binaural spatial features was shown to recognize sound events better than using just the monaural features.

Feature combination	Indoor							Outdoor		
	Basketball	Bus	Hallway	Office	Car	Restaurant	Shop	Beach	Street	Track and Field
mel-$monaural$	79.7	52.6	59.1	81.8	78.2	80.7	62.4	56.5	60.3	70.1
mel	82.2	56.5	66.6	83.3	81.5	83.1	63.3	59.5	66.0	70.2
mel + $TDOA$	82.8	58.7	66.0	80.8	79.2	81.2	64.7	**60.9**	**66.9**	68.8
mel + GCC-$PHAT$	81.9	58.9	65.3	80.0	81.2	81.3	65.3	60.4	66.3	**72.6**
mel + dom-$freq$	**83.7**	**60.5**	**67.8**	**84.6**	80.8	81.8	64.6	60.7	66.6	67.6
mel + ACR	82.9	58.6	63.8	83.6	**83.4**	**82.3**	**65.5**	60.4	65.8	69.0
mel + $TDOA$ + dom-$freq$	83.0	59.4	67.5	83.9	78.6	79.9	65.1	60.5	65.0	70.0
mel + GCC-$PHAT$ + ACR	82.8	59.1	66.8	82.2	79.4	80.4	64.8	60.4	66.1	68.5

Table 2. Context wise F-scores for TUT-SED 2009 dataset.

6. REFERENCES

[1] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015.

[2] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time-frequency audio features," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," in *ACM Computing Surveys (CSUR)*, 2016.

[4] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[5] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2008.

[6] A. Temko, C. Nadeu, and J. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07," in *Springer-Verlag, Berlin*, 2008.

[7] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[8] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[9] J. W. Strutt, "On our perception of sound direction," in *Philosophical Magazine*, 1907.

[10] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[11] C. Knapp and C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976.

[12] J. O. Smith, *Sinusoidal Peak Interpolation, in Spectral Audio Signal Processing*, accessed 23.06.2016, online book, 2011 edition. [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/Sinusoidal_Peak_Interpolation.htm

[13] A. S. Bregman, "Auditory scene analysis: The perceptual organization of sound," in *MIT Press*, 1990.

[14] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in *IEEE/ACM TASLP Special Issue on Sound Scene and Event Analysis*, 2017, accepted for publication.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," in *Journal of Machine Learning Research (JMLR)*, 2014.

[17] P. J. Werbos, "Backpropagation through time: what it does and how to do it," in *Proceedings of the IEEE*, 1990.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv:1412.6980 [cs.LG]*, 2014.

[19] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in *European Signal Processing Conference (EUSIPCO)*, 2010.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *European Signal Processing Conference (EUSIPCO)*, 2016.

[21] "Detection and classification of acoustic scenes and events (DCASE)," 2016. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio

[22] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, 2016.

Publication III

Sharath Adavanne, Archontis Politis, Tuomas Virtanen, "Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features," *International Joint Conference on Neural Networks (IJCNN).* Rio de Janeiro, Brazil, pp. 1-7, July 2018.

Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features

Sharath Adavanne[1], Archontis Politis[2], Tuomas Virtanen[1]
[1]Laboratory of Signal Processing, Tampere University of Technology, Finland
Email: firstname.lastname@tut.fi
[2]Department of Signal Processing and Acoustics, Aalto University, Finland
Email: archontis.politis@aalto.fi

Abstract—**In this paper, we propose a stacked convolutional and recurrent neural network (CRNN) with a 3D convolutional neural network (CNN) in the first layer for the multichannel sound event detection (SED) task. The 3D CNN enables the network to simultaneously learn the inter- and intra-channel features from the input multichannel audio. In order to evaluate the proposed method, multichannel audio datasets with different number of overlapping sound sources are synthesized. Each of this dataset has a four-channel first-order Ambisonic, binaural, and single-channel versions, on which the performance of SED using the proposed method are compared to study the potential of SED using multichannel audio. A similar study is also done with the binaural and single-channel versions of the real-life recording TUT-SED 2017 development dataset. The proposed method learns to recognize overlapping sound events from multichannel features faster and performs better SED with a fewer number of training epochs. The results show that on using multichannel Ambisonic audio in place of single-channel audio we improve the overall F-score by 7.5 %, overall error rate by 10 % and recognize 15.6 % more sound events in time frames with four overlapping sound sources.**

I. Introduction

Sound event detection (SED) is the task of recognizing the sound events and their respective temporal start and end time in an audio recording. Sound events in real life do not always occur in isolation but tend to considerably overlap with each other. Recognizing such overlapping sound events is referred as polyphonic SED. Applications of such polyphonic SED are numerous. Recognizing sound events like alarm and glass breaking can be used for surveillance [1], [2]. Automatic detection of road accidents can ensure quick intervention of emergency teams [3]. Environmental sound event detection can be used for monitoring biodiversity [4], [5], [6]. Further, SED can be used for automatically annotating audio datasets, and the sound events recognized can be used as a query for similar content retrieval.

Polyphonic SED using single-channel audio has been studied extensively. Different approaches have been proposed using supervised classifiers like Gaussian mixture model - hidden Markov model [7], fully-connected networks [8], convolutional neural networks (CNN) [9], [10], and recurrent neural networks (RNN) [11], [12], [13]. More recently, the state of the art method for polyphonic SED was proposed in [14], where the log mel-band energy feature was used with a convolutional recurrent neural network (CRNN) architecture.

Recognizing overlapping sound events using a single-channel audio is a challenging task. These overlapping sound events can potentially be recognized better with multichannel audio. One of the first methods to use multichannel audio for SED was proposed in [15], which performed SED on each of the audio channels separately and the combined likelihoods across channels were used for the final prediction. More recently the state of the art CRNN network for single-channel SED [14] was extended for multichannel features and multiple feature classes in [16]. It was shown that the performance of SED improves on using the binaural audio instead of the single-channel audio version of the same dataset. In this regard, [16] also proposed binaural audio features exploiting the inter-aural intensity and time differences. In the network proposed by [16] the CNNs were used as feature extractors that learned just the intra-channel information from the input multichannel audio features, while the RNNs which followed the CNNs were learning the inter-channel information. In this paper, we propose to learn both the inter- and intra-channel information within the CNN layer. We implement this by using a 3D CNN [17] as the first layer of the network. This enables the method to learn both inter- and intra-channel information from the input multichannel audio within the CNN layers for no additional parameters in comparison to [16].

The hardware devices for smart homes, virtual reality content creation, modern hearing aids and surveillance sensors have more than one microphone in them. By using all the multichannel audio available from these devices we can potentially improve the polyphonic SED, and this improvement can additionally enhance the overall performance of these devices. Although [16] showed that using binaural audio in place of single-channel audio improves the performance of SED, there is no other conclusive work that studies the potential of SED

The research leading to these results has received funding from the European Research Council under the European Unions H2020 Framework Programme through ERC Grant Agreement 637422 EVERYSOUND. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

with more than two-channel of audio. Besides, in order to carry out such a study, there are no publicly available data. Moreover, collecting and annotating such a dataset for SED is a tedious, expensive and time-consuming task. In order to assess the necessity of collecting such a dataset, in this paper, we synthesize three multichannel audio datasets with up to one, up to three and up to six temporally overlapping sound sources. The multichannel audio in each of the datasets is a four-channel first-order Ambisonic (FOA) audio. Additionally, we perform binauralization with real head related transfer function (HRTF) to obtain binaural version from the FOA audio, and further used the omnidirectional channel of FOA as the single-channel version. Experiments are carried out on these datasets to understand the extent of improvement we can achieve by using multichannel audio over the current state of the art SED methods using single-channel and binaural audio. Based on the results obtained we can decide to invest in the collection of real-life multichannel dataset. Furthermore, in order to compare the consistency of results obtained with the synthetic dataset, we perform similar experiments on the real-life recordings TUT-SED 2017 dataset [18], that consists of only the single-channel and binaural audio.

The paper is organized as follows: Section II describes the feature extraction details and the proposed neural network. The datasets used, metric for evaluation, the baseline method and the evaluation procedure are explained in Section III. Finally, the results and discussion are presented in Section IV.

II. Method

The proposed multichannel SED method is shown in Figure 1. The input to the method is either a single-channel or multichannel audio. For single-channel audio input, we use just the log mel-band energy feature. In the case of multichannel input, the log mel-band energy feature is extracted in each of the channels; additionally, generalized cross-correlation with phase transform (GCC-PHAT) [19] feature is extracted between each channel pair of the multichannel audio. These audio features are fed to a multichannel neural network architecture that maps them to the activities of the sound event classes in the dataset. The output of the neural network is in the continuous range of $[0, 1]$ for each of the sound event classes and corresponds to the probability of the particular sound class being active in the frame. This continuous range output is further thresholded to obtain the final binary decision of the sound event class being active or absent in each frame. In general, the proposed method takes a sequence of frame-wise audio features as the input and predicts the activity of the target sound event classes for each of the input frames. The detailed description of feature extraction and the neural network is presented below.

A. Feature extraction

1) Log mel-band energy: Previously, single-channel SED methods have been using log mel-band energies (mbe) and have shown to be effective for the task [8], [11], [14], [16], [20], [21]. Additionally, in the case of SED with binaural audio, mbe extracted from the two channels was proposed in [11]. This was motivated from the inter-aural intensity difference (IID) used by the human auditory system to localize and recognize overlapping sound events. A neural network that is capable of performing linear operations (which includes the difference operation) can obtain information similar to the IID from the binaural mbe. More recently, the binaural mbe was shown to improve the performance of SED even on larger binaural datasets [16]. Motivated from this, we continue to use mbe feature extracted from all the input channels in this paper.

In case of a single-channel audio input, we extract mbe in 40 ms windows with 50% overlap and refer to it as mbe-$mono$. We use 40 mel-bands in the frequency range of 0-22050 Hz. For multichannel audio we extract mbe in each of the channels, and refer to it as mbe-bin for binaural and mbe-$ambi$ for four-channel FOA audio. For a sequence length of T frames, the mbe feature has a general dimension of $T \times 40 \times C$, where C is the number of channels, $C = 1$ for mbe-$mono$, $C = 2$ for mbe-bin and $C = 4$ for mbe-$ambi$.

2) Generalized cross correlation with phase transform: In the case of binaural audio, [16] proposed to represent similar information as the inter-aural time difference (ITD) in humans using the generalized cross correlation with phase transform (gcc). It was shown that the SED methods can benefit with gcc for overlapping sound events. Motivated from this, we continue to use gcc in this paper. Similar to [16], we extract gcc in three resolutions, 120, 240, and 480 ms as

$$R(\Delta_{12}, t) = \sum_{k=0}^{K-1} \frac{X_1(k,t) \cdot X_2^*(k,t)}{|X_1(k,t)||X_2(k,t)|} e^{\frac{i2\pi k \Delta_{12}}{N}}, \qquad (1)$$

where, X_1 and X_2 are the FFT coefficients of the two-channels between which the gcc is calculated. $X_1(k, t)$ is the coefficient at time frame t and kth frequency bin, of the total K bins. gcc per frame given by $R(\Delta_{12}, t)$ is extracted for delays Δ_{12} in the range $[-\tau_{\text{max}}, \tau_{\text{max}}]$, where τ_{max} is the maximum sample delay for a sound wave to travel between the pair of microphones recording audio. In order to have a factorisable feature length for max pooling in the neural network, 60 gcc values are chosen in the range $\Delta_{12} \in [-29, 30]$ lag for each of the three multi-resolution. For a sequence length of T frames, the gcc feature is of the general dimension $T \times 60 \times 3\binom{C}{2}$, where $\binom{C}{2}$ is the number of possible pair-of-two combinations for the C channels of audio (denoted in Figure 1 as C_2) and 3 is the number of resolutions in which gcc was extracted. In the case of binaural audio (gcc-bin) this results in $T \times 60 \times 3$ and for Ambisonic audio (gcc-$ambi$) this amounts to $T \times 60 \times 18$.

B. Neural network

The input to the proposed method is $T \times 40 \times C$ dimensional mbe and $T \times 60 \times 3\binom{C}{2}$ dimensional gcc features as shown in Figure 1. Based on the task of single or multichannel SED, the network is fed with the respective feature sequence.

Separate CNN branches are used to learn local shift-invariant features from each of the input features mbe and gcc. The first CNN layer in each CNN branch consists of a

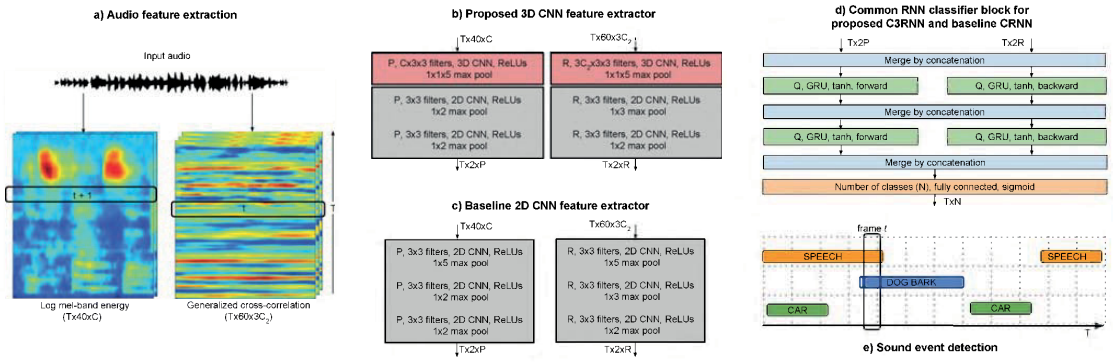


Fig. 1. The proposed C3RNN (a+b+d+e) and baseline CRNN (a+c+d+e) stacked convolutional and recurrent neural network architectures for multichannel polyphonic sound event detection.

3D CNN, i.e., convolution over volumes. The receptive filters of 3D CNNs are of the size $D \times 3 \times 3$ size, where $D = C$ for mbe and $D = 3\binom{C}{2}$ for gcc feature. This joint learning of features along channel-time-frequency enables the network to learn both inter- and intra-channel features simultaneously within the first layer. The 3D CNN is followed by a sequence of 2D CNN layers with receptive filters of size 3×3. The output activation from both the CNN layers is padded with zeros to keep the dimension of the output the same as the input. Batch normalization [22] and max-pooling is performed after every layer of CNN along frequency axis to reduce the final dimension to $T \times 2 \times P$ for mbe and $T \times 2 \times R$ for gcc, where P and R are the number of filters in the final layer of CNN in respective CNN branches. The CNN activations from the two branches are concatenated along feature axis and are fed to layers of bi-directional gated recurrent units (GRU), to learn long-term temporal activity patterns. This is followed by a layer of time-distributed fully-connected (dense) network. The final prediction layer has as many sigmoid units as the number of sound event labels in the dataset. We refer to this network as C3RNN in future.

The training is performed for 1000 epochs using Adam [23] optimizer, and binary cross-entropy loss between the reference sound class activities and the predicted ones. Dropout [24] is used as a regularizer after every layer of the neural network to make it robust to unseen data. Early stopping is used to stop overfitting the network to training data. A threshold of 0.5 is used to obtain the binary decision from the sigmoid activations in the final prediction layer. Training is stopped if the error rate (see Section III-B) on the test split does not improve for 100 epochs. The neural network implementation was done using the PyTorch [25] library.

III. Evaluation

A. Dataset

We evaluate the proposed C3RNN with four different datasets, one real-life audio TUT-SED 2017 Development dataset [18] and three synthetic datasets. The recordings of TUT-SED 2017 are binaural. In order to assess the performance of SED for more than two channels of audio we propose to use the synthetic datasets.

1) TUT-SED 2017 Development dataset: This dataset was recorded in the street context using a binaural in-ear microphone at 24 bit and 44.1 kHz sampling rate. Each of the recordings is of the length 3-5 minutes, amounting to a total length of 70 minutes. This dataset consists of manual annotations for sound event classes such as brakes squeaking, car, children, large vehicle, people speaking, and people walking. The dataset defines four-folds of training and testing splits for benchmarking. Further details of the dataset are given in [18]. Since the dataset has only two channels, we do not have mbe-$ambi$ and gcc-$ambi$ features for this dataset. The single-channel version is obtained by taking the mean of the binaural channels.

2) Synthetic dataset: In order to assess the performance of SED in presence of more than two channels of audio, we generate synthetic datasets using the method proposed in [26]. Three separate anechoic multichannel datasets with a) no temporally overlapping sources ($O1$), b) maximum three overlapping sources ($O3$), and c) maximum six overlapping sources ($O6$) are synthesized. For each dataset, three sets of training and test split were generated, each with 500 and 100 recordings respectively. Every recording is of length 30 seconds and sampled at 44100 Hz. The dataset consists of only stationary point sources. Point sources are sound events which can be associated with a single spatial coordinate in the space, for example, a person speaking, or a phone ringing. Diffuse sources like ambient noise, wind breeze, etc. do not have a specific spatial coordinate and are therefore more difficult to synthesize spatially, hence we do not use them in this study.

The audio recordings synthesized were of first-order Ambisonic (FOA) format. This is a commonly used format for spatial audio, especially in the virtual reality domain[1]. The

[1]https://developers.google.com/vr/concepts/spatial-audio

FOA consists of four channels of audio, commonly referred as W, X, Y, and Z channels, where, X, Y and Z channels represents the directive pressure-gradient recordings along the X, Y and Z axes of the Cartesian coordinate system respectively. The W channel corresponds to an omnidirectional microphone recording. In this paper, we use the W channel for our single-channel SED studies and all the four channels (W, X, Y, and Z) for our four-channel SED studies. We further perform binauralization of the four-channel audio using real head-related transfer functions (HRTF) to obtain the binaural version and used them for the binaural SED studies. The HRTFs were measured from one of the authors on a dense grid of directions under anechoic conditions, as detailed in [27]. For an overview on HRTF measurement techniques, and simulation of spatial sound scenes based on them, such as in this work, the reader is referred to [28].

In order to synthesize these datasets, we use the isolated sound events from the DCASE 2016 task 2 [29]. This dataset consists of 11 sound event classes, with 20 examples each. The sound event classes include speech, cough, door slam, laughter, phone, knock. We chose 16 examples from each class randomly for training and four for the testing split. In order to synthesize a recording, each sound example was randomly associated with a spatial coordinate such that two temporally overlapping examples do not have the same spatial coordinate. Further, the magnitude of the sound examples was varied randomly to give the effect of varying distance from the microphone. Details of the synthesis procedure are given in [26].

B. Metric

The proposed SED method is evaluated using the polyphonic SED metrics proposed in [30]. Particularly we use segment wise error rate (ER) and F-score calculated in one-second length segments. The F-score is calculated as

$$F = \frac{2 \cdot \sum_{k=1}^{K} TP(k)}{2 \cdot \sum_{k=1}^{K} TP(k) + \sum_{k=1}^{K} FP(k) + \sum_{k=1}^{K} FN(k)}, \tag{2}$$

where for each one-second segment k, $TP(k)$ is the number of true positives i.e., the number of sound event labels active in both predictions and ground truth. $FP(k)$ is the number of false positives i.e., the number of sound event labels active in predictions but inactive in ground truth. $FN(k)$ is the number of false negatives i.e., the number of sound event labels active in the ground truth but inactive in the predictions.

The error rate is measured as

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)}, \tag{3}$$

where $N(k)$ is the total number of active sound events in the ground truth of segment k. The number of substitutions $S(k)$, deletions $D(k)$ and insertions $I(k)$ is measured using the following equations for each of the K one second segments:

$$S(k) = \min(FN(k), FP(k)) \tag{4}$$

$$D(k) = \max(0, FN(k) - FP(k)) \tag{5}$$

$$I(k) = \max(0, FP(k) - FN(k)) \tag{6}$$

According to the Equations (2) and (3), for an ideal SED method, ER is zero and F-score is one. In this paper, we report the F-score in percentage and hence the ideal F-score will be 100 %.

C. Baseline

The proposed C3RNN is compared with the existing state of the art multichannel audio SED method proposed in [16]. Similar to the proposed C3RNN, the baseline method can perform SED with single-channel, binaural and multichannel audio. Previously, its performance has only been tested with single-channel and binaural audio. This method won [31] the recently concluded IEEE Audio and Acoustic Signal Processing research challenge – DCASE 2017 Task 3 for real life sound event detection [18]. In particular, it secured the first two positions among the 34 submitted methods. The first position was obtained with the *mbe-mono* audio feature and a close second position with *mbe-bin*. This proves that the method is well suited for both single-channel and binaural SED baselines.

The baseline method shown in Figure 1 is also based on a stacked convolutional and recurrent neural network (CRNN). In comparison to the proposed C3RNN, the method does not employ a 3D CNN, thus its CNN only learns intra-channel information, while the RNNs learn the inter-channel information. The rest of the inputs and the outputs of the baseline CRNN and proposed C3RNN are similar. In this paper, we consider both the CRNN with single-channel and binaural audio features as the baselines, and further report the performance of CRNN with multichannel Ambisonic audio along with the C3RNN performance.

D. Evaluation procedure

In order to evaluate the performance of the proposed C3RNN with respect to the baseline CRNN on multichannel dataset, the two methods were trained individually using the single-channel, binaural, and Ambisonic audio features of the synthetic dataset and the single-channel, and binaural audio features of TUT-SED 2017 development dataset. We perform a hyper-parameter search on each of the dataset-feature combinations individually and assess the performance of SED using multichannel audio using the ER and F-scores on the test splits. The metric scores reported are the mean of five separate runs on the cross-validation splits.

In order to study the individual contribution of *gcc* for the SED task, we performed an experiment of estimating the number of sound sources in every time frame using just the *gcc* feature. The usage of *gcc* for SED task was motivated from the idea that the relative time difference of arrival of two overlapping sound sources will be different, and this will

TABLE I
THE EVALUATION METRIC SCORES FOR THE SOUND EVENT DETECTION TASK USING THE PROPOSED C3RNN AND BASELINE CRNN WITH *mbe* AND *gcc* AUDIO FEATURE FOR DIFFERENT OVERLAPPING SOUND EVENTS DATASETS.

	O1		*O3*		*O6*	
C3RNN	ER	F	ER	F	ER	F
mbe-gcc-ambi	0.11	92.2	0.18	82.5	0.17	84.1
mbe-gcc-bin	0.12	91.6	0.20	79.8	0.24	77.2
mbe-ambi	**0.09**	93.7	**0.16**	**83.8**	**0.16**	**85.4**
mbe-bin	0.10	**93.8**	0.18	81.8	0.22	78.5
mbe-mono	0.10	91.9	0.17	81.8	0.26	77.9
CRNN	ER	F	ER	F	ER	F
mbe-gcc-ambi	0.11	91.1	0.19	81.6	0.19	83.5
mbe-gcc-bin	0.12	92.3	0.21	78.8	0.26	79.0
mbe-ambi	**0.10**	92.8	**0.18**	**82.5**	**0.17**	**83.7**
mbe-bin	0.11	**93.6**	0.19	79.3	0.23	79.5
mbe-mono	0.12	91.9	0.18	80.6	0.28	78.3

be highlighted in the gcc feature. In the proposed experiment of identifying the number of active sources, using just gcc feature should have better accuracy than using only the mbe feature. This would mean that the mbe based SED methods will additionally benefit from using gcc.

We trained the proposed C3RNN with just gcc feature as input and the number of active sound sources as the output. Similar training was done using just mbe feature as input. When using a single feature in the proposed C3RNN method, for example, the mbe feature, the CNN feature extractor branch for gcc is removed, and only the CNN feature extractor branch for mbe is used. Separate hyper-parameter search was done for the individual features randomly [32], and the best configurations for both gcc-$ambi$ and mbe-$ambi$ features for synthetic dataset $O6$ had around 270 k trainable weights. Unlike the SED task which is a multi-label classification task (more than one sound event can be active in a given time frame), this experiment of estimating the number of sources is a multi-class classification task (mutually exclusive classes). Hence, for this experiment alone the output sigmoid activation was replaced with softmax, and the categorical cross entropy loss was used.

IV. RESULTS AND DISCUSSION

A hyperparameter search was carried out with the proposed C3RNN and baseline CRNN for each combination of the dataset (synthetic $O1$, $O3$ and $O6$, and TUT-SED 2017 development dataset) and audio feature (mbe-$mono$, mbe-bin, mbe-$ambi$, mbe-gcc-bin, and mbe-gcc-$ambi$). In general, the hyperparameters remained the same for a given dataset, independent of the feature used. A sequence length of 128 frames, batch size of 32 and dropout of 0.35 gave the best results for all the feature and synthetic dataset combinations. For the TUT-SED 2017 dataset, the best results were obtained using a sequence length of 256 frames, a batch size of 128 and dropout of 0.2. A learning rate of 1×10^{-4} gave the best results across datasets and audio features. The performance was not affected much by the exact number of CNN filters or GRU units. Across datasets, different number of CNN filters and GRU units were seen to give good evaluation metric scores. In the case of the synthetic $O1$ dataset, the optimal number of CNN filters for mbe in each layer (P in Figure 1) was 8, for gcc in each layer (R in Figure 1) it was 16 and the GRU units in each layer (Q in Figure 1) was 8. Similarly $P = Q = 16$ and $R = 32$ for synthetic $O3$, $P = Q = 32$ and $R = 64$ for synthetic $O6$, and $P = Q = R = 64$ for TUT-SED 2017 dataset. This correlation of increasing number of CNN filters and GRU units with increasing number of overlapping sound events in datasets shows that bigger neural networks are required for recognizing highly overlapped sound events.

The evaluation results for the proposed C3RNN using single-channel, binaural and Ambisonic audio mbe and gcc features for different polyphonic datasets are presented in Table I. Analyzing the performance of mbe only features first, we see that for no polyphony ($O1$) the ER and F-scores are comparable for single (mbe-$mono$) and multichannel (mbe-bin and mbe-$ambi$). With the increase in polyphony ($O3$ and $O6$), the ER and F scores of binaural and multichannel SED improves over the single-channel. Particularly, this improvement is significant for the dataset with highly overlapping sound events ($O6$). Concretely, using mbe-$ambi$ instead of mbe-$mono$ on $O6$ dataset gives 7.5% improvement in F-score and 10% in ER. A similar trend is observed using baseline CRNN for mbe only feature and the results are comparable to proposed C3RNN. For the mbe-$mono$ feature the baseline CRNN and proposed C3RNN should ideally have the exact same scores across datasets since there is no additional inter-channel information for the C3RNN to learn from. The deviations seen in the metric scores are from the random initializations of the network even after averaging the scores from five separate runs on the cross-validation data. The actual improvement of using the proposed C3RNN over the baseline CRNN is achieved in training speed. As shown in Figure 2, C3RNN achieves better error rate with lower

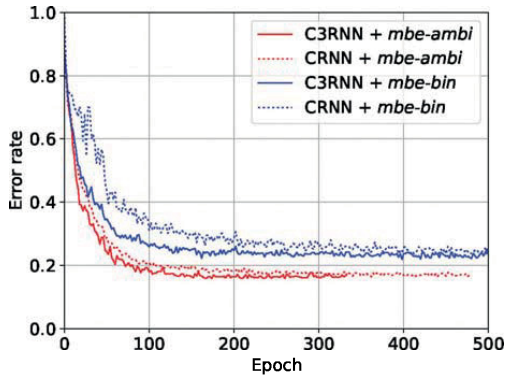


Fig. 2. The learning curve for the proposed C3RNN and baseline CRNN methods, for ambisonic (mbe-$ambi$) and binaural (mbe-bin) features of the synthetic $O6$ dataset. The proposed C3RNN achieves better error rate with a lower number of epochs, for both mbe-$ambi$ and mbe-bin features.

TABLE II
FRAMEWISE ACCURACY (IN %) OF RECOGNIZING THE CORRECT NUMBER OF SOUND EVENTS IN THE SYNTHETIC $O6$ DATASET.

	Number of overlapping sound events							
C3RNN	0	1	2	3	4	5	6	Avg.
gcc-ambi	90.7	46.0	38.0	34.4	29.5	10.4	0.0	35.6
mbe-ambi	**92.7**	**66.9**	**56.3**	**47.7**	**34.7**	**16.3**	0.3	**45.0**
mbe-bin	90.4	58.0	46.3	39.8	27.8	12.7	**0.6**	39.4
mbe-mono	89.9	60.8	48.1	35.2	19.1	8.6	0.1	37.4
CRNN								
mbe-ambi	**93.5**	**66.4**	**56.5**	**47.3**	**32.4**	**15.7**	**0.5**	**44.6**
mbe-bin	92.8	60.6	47.7	42.9	29.1	12.3	0.2	40.8
mbe-mono	90.8	59.6	49.9	34.1	18.4	9.7	0.4	37.6

number of epochs for both *mbe-bin* and *mbe-ambi* features. The proposed C3RNN achieves this with exactly the same number of weights as the baseline CRNN, but with different convolution connections in CNN feature extraction layer.

Another observation from Table I for *mbe-mono* feature and across methods is that the performance of SED drops with a higher number of overlapping sound events. Using multichannel features, especially *mbe-ambi*, the performance is comparable for up to three ($O3$) and six ($O6$) overlapping sound events datasets. In general, the *mbe-ambi* is seen to perform better SED than the *mbe-bin*, which in turn performs better than *mbe-mono*. Additionally, the SED performance is seen to significantly improve with multichannel audio for sound scenes with highly overlapping sound events. This shows that using additional audio channel information definitely helps in more reliable and robust SED.

Table I also reports the performance of using *mbe* and *gcc* features together (*mbe-gcc*). Since *gcc* can only be extracted for more than one channel of audio, it reports results only for the binaural and Ambisonic versions of audio. In comparison to its respective *mbe* only features, the evaluation metric scores are either comparable or worse for both C3RNN and baseline CRNN methods. To investigate this, and understand if using *gcc* feature provides additional information to *mbe*, the experiment of estimating the number of sound sources per frame was carried out. An average accuracy of 35.6 % was obtained in estimating the number of sound sources per frame using just *gcc-ambi* (see Table II), while *mbe-ambi* alone gave 45.0 %. Similar results were obtained using binaural audio on the synthetic $O6$ dataset and the TUT-SED 2017 dataset (see Table III). Although the usage of *gcc* feature in addition to *mbe* has been shown to be helpful for other binaural SED datasets [11], [16], the present results in Table I show that it does not provide any additional information for the SED datasets studied in this paper. The dominance of the *mbe* features could be explained by the strong head shadowing effects at different source directions in binaural recordings and the spatial coincidence of Ambisonic recordings that encodes spatial information based only on inter-channel level differences. This dominance of *mbe* feature may not hold for audio formats which rely on phase- or time-differences to encode directional information, with insignificant level differences.

TABLE III
FRAMEWISE ACCURACY (IN %) OF RECOGNIZING THE CORRECT NUMBER OF SOUND EVENTS IN THE TUT-SED 2017 DATASET.

	Number of overlapping sources				
C3RNN	0	1	2	3	Avg.
mbe-bin	**70.1**	**70.2**	**73.1**	**16.4**	**57.5**
gcc-bin	62.2	64.6	39.4	2.0	42.1

TABLE IV
THE EVALUATION METRIC SCORES FOR THE SOUND EVENT DETECTION TASK USING THE PROPOSED C3RNN AND BASELINE CRNN FOR THE TUT-SED 2017 DATASET.

	C3RNN		CRNN	
	ER	F	ER	F
mbe-bin	**0.35**	**67.5**	**0.37**	**64.8**
mbe-mono	0.38	64.1	0.39	63.3

Audio captured with linear arrays or spaced omnidirectional microphones are examples of such audio formats and may benefit significantly from *gcc* features instead of the level differences captured in *mbe*.

Among the audio features in Table II, we see that using multichannel features, especially *mbe-ambi* significantly improves the accuracy of estimating overlapping sound events in comparison to single-channel *mbe-mono* feature. In numbers, using *mbe-ambi* instead of *mbe-mono* in the proposed C3RNN method improves the performance of detection of three overlapping sources by 12.5 % and four overlapping sources by 15.6 %. This proves that using multichannel audio for SED helps recognize overlapping sound events better than single-channel audio.

The evaluation metric scores for the real-life recordings TUT-SED 2017 dataset is presented in Table IV. The results are consistent with the results obtained with synthetic datasets. The performances of C3RNN and CRNN are comparable, and the multichannel feature *mbe-bin* achieves better SED than single-channel *mbe-mono* feature. Additionally, the error rate obtained using the proposed C3RNN and *mbe-bin* feature beats the current top score of 0.50 [31] on this benchmarking dataset.

V. CONCLUSION

In this paper, we proposed a stacked convolutional and recurrent neural network with inter- and intra-channel convolutions in the first layer (C3RNN) for the multichannel sound event detection (SED) task. The inter- and intra-channel convolutions were implemented using a 3D convolutional neural network (CNN) layer. It was shown that the proposed C3RNN method learns to recognize overlapping sound events from multichannel features faster than the state of the art baseline multichannel SED method with exactly the same number of parameters, and further performs better SED with fewer number of training epochs. In the multichannel SED task, for the SED datasets used in the paper, it was shown that the generalized cross-correlation with phase transform feature was not providing any additional information to the standard multichannel log mel-band energy feature.

Additionally, we proposed to assess the performance of using multichannel audio for polyphonic SED. The study was carried out on four datasets- a) one real-life recording TUT-SED 2017 development dataset, and three synthetic datasets with b) no temporally overlapping, c) up to three temporally overlapping and d) up to six temporally overlapping sound events. Each of the recordings in the synthetic datasets was in three formats, four-channel first-order Ambisonics, binaural and single-channel, whereas the real-life dataset had just the binaural and single-channel audio versions. We performed SED individually on these datasets using the proposed C3RNN. In comparison to using a single-channel, we observed that by using multichannel audio, the overall F-score improved by 7.5 %, overall ER improved by 10 % and 15.6 % more sound events were recognized in time frames with four overlapping sound events. In conclusion, multichannel audio definitely improves the SED performance over using single-channel audio; and the collection of such a real-life multichannel audio dataset is worth the effort.

REFERENCES

[1] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," in *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, 2016.

[2] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," in *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 17, no. 1, 2015.

[4] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time-frequency audio features," in *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 17, no. 6, 2009.

[5] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack, "Estimating animal population density using passive acoustics," in *Biological reviews of the Cambridge Philosophical Society*, vol. 88, no. 2, 2012.

[6] B. J. Furnas and R. L. Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," in *Journal of Wildlife Management*, vol. 79, no. 2, 2014.

[7] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *European Signal Processing Conference (EUSIPCO)*, 2010.

[8] E. Çakır, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi-label deep neural networks," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2015.

[9] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[10] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," in *INTERSPEECH*, 2016.

[11] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[12] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, and K. Takeda, "Duration-controlled LSTM for polyphonic sound event detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 11, 2017.

[13] M. Zöhrer and F. Pernkopf, "Virtual adversarial training and data augmentation for acoustic event detection with gated recurrent neural networks," in *INTERSPEECH*, 2017.

[14] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in *IEEE/ACM Transacation on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 6, 2017.

[15] A. Temko, C. Nadeu, and J. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07," in *Springer-Verlag, Berlin*, 2008.

[16] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[18] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE2017 challenge setup: Tasks, datasets and baseline system," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[19] C. Knapp and C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, vol. 24, no. 4, 1976.

[20] I.-Y. Jeong, S. Lee, Y. Han, and K. Lee, "Audio event detection using multiple-input convolutional neural network," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[21] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv:1412.6980 [cs.LG]*, 2014.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," in *Journal of Machine Learning Research (JMLR)*, vol. 15, no. 1, 2014.

[25] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Neural Information Processing Systems (NIPS) Workshop on Autodiff*, 2017.

[26] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference (EUSIPCO)*, 2018, submitted for review. [Online]. Available: https://arxiv.org/abs/1710.10059

[27] J. G. Bolaños and V. Pulkki, "HRIR database with measured actual source direction data," in *Audio Engineering Society Convention*, 2012.

[28] B. Xie, *Head-related transfer function and virtual auditory display*, 2nd ed. Plantation, FL: J. Ross Publishing, 2013.

[29] E. Benetos, M. Lagrange, and G. Lafay, http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-synthetic-audio#audio-dataset, accessed on 17.01.2018.

[30] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, vol. 6, no. 6, 2016.

[31] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[32] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," in *Journal of Machine Learning Research*, vol. 13, no. 1, 2012.

Publication IV

Sharath Adavanne, Archontis Politis, Tuomas Virtanen, "Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network," *European Signal Processing Conference (EUSIPCO)*. Rome, Italy, pp. 1462-1466, September 2018.

Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network

Sharath Adavanne*[1], Archontis Politis*[2], Tuomas Virtanen[1]
[1]Laboratory of Signal Processing, Tampere University of Technology, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Finland

***Abstract*—This paper proposes a deep neural network for estimating the directions of arrival (DOA) of multiple sound sources. The proposed stacked convolutional and recurrent neural network (DOAnet) generates a spatial pseudo-spectrum (SPS) along with the DOA estimates in both azimuth and elevation. We avoid any explicit feature extraction step by using the magnitudes and phases of the spectrograms of all the channels as input to the network. The proposed DOAnet is evaluated by estimating the DOAs of multiple concurrently present sources in anechoic, matched and unmatched reverberant conditions. The results show that the proposed DOAnet is capable of estimating the number of sources and their respective DOAs with good precision and generate SPS with high signal-to-noise ratio.**

I. Introduction

Direction of arrival (DOA) estimation is the task of identifying the relative position of the sound sources with respect to the microphone. DOA estimation is a fundamental operation in microphone array processing and forms an integral part of speech enhancement [1], multichannel sound source separation [2] and spatial audio coding [3]. Popular approaches to DOA estimation are based on time-delay-of-arrival (TDOA) [4], the steered-response-power (SRP) [5], or on subspace methods such as multiple signal classification (MUSIC) [6] and the estimation of signal parameters via rotational invariance technique (ESPRIT) [7].

The aforementioned methods differ from each other in terms of algorithmic complexity, and their suitability to various arrays and sound scenarios. MUSIC specifically is very generic with regards to array geometry, directional properties and can handle multiple simultaneously active narrowband sources. On the other hand, MUSIC and subspace methods in general, require a good estimate of the number of active sources, which are often unavailable or difficult to obtain. Furthermore, MUSIC can suffer at low signal to noise ratio (SNR) and in reverberant scenarios [8]. In this paper, we propose to overcome the above shortcomings with a deep neural network (DNN) method, referred to as DOAnet, that learns the number of sources from the input data, generates high precision DOA estimates and is robust to reverberation. The proposed DOAnet also generates a spatial acoustic activity map similar to the MUSIC pseudo-spectrum (SPS) as an intermediate output. The SPS has numerous applications that rely on a directional map of acoustic activity such as soundfield visualizations [9], and room acoustics analysis [10]. In comparison, the proposed DOAnet outputs the SPS and DOA's of multiple overlapping sources similar to any popular DOA estimators like MUSIC, ESPRIT or SRP without requiring the critical information of the number of active sound sources. A successful implementation of this will enable the integration of such DNN methods to higher-level learning based end-to-end sound analysis and detection systems.

Recently, several DNN-based approaches have been proposed for DOA estimation [11], [12], [13], [14], [15], [16]. There are six significant differences between them and the proposed method: a) All the aforementioned works focused on azimuth estimation, with the exception of [15] where the 2-D Cartesian coordinates of sound sources in a room were predicted, and [11] trained separate networks for azimuth and elevation estimation. In contrast, we demonstrate the estimation of both azimuth and elevation for the DOA by sampling the unit sphere uniformly and predicting the probability of sound source at each direction. b) The past works focused on the estimation of a single DOA at every time frame, with the exception of [13] where localization of azimuth for up to two sources simultaneously was proposed. On the other hand, the proposed DOAnet does not algorithmically limit the number of directions to be estimated, i.e., with a higher number of audio channels input, the DOAnet can potentially estimate a larger number of sound events.

c) Past works were evaluated with different array geometries making comparison difficult. Although the DOAnet can be applied to any array geometry, we evaluate the method using real spherical harmonic input signals, which is an emerging popular spatial audio format under the name Ambisonics. Microphone signals from various arrays, such as spherical, circular, planar or volumetric, can be transformed to Ambisonic signals by an appropriate transform [17], resulting in a common representation of the 3-D sound recording. Although the DOAnet is scalable to higher-order Ambisonics, in this paper we evaluate it using the compact four-channel first-order Ambisonics (FOA).

d) Regarding classifiers, earlier methods have used fully

*Equally contributing authors in this paper. The research leading to these results has received funding from the European Research Council under the European Unions H2020 Framework Programme through ERC Grant Agreement 637422 EVERYSOUND. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources

connected (FC) neural networks [11], [12], [13], [14], [15] and convolutional neural networks (CNN) [16]. In this work, along with the CNNs we use recurrent neural network (RNN) layers. The usage of RNN allows the network to learn long-term temporal information. Such an architecture is referred to as a convolutional recurrent neural network (CRNN) in literature and is the state-of-the-art method in many single- [18], [19] and multichannel [20], [21] audio tasks. e) Previous methods used inter-channel features such as generalized cross-correlation with phase transform (GCC-PHAT) [15], [12], eigen-decomposition of the spatial covariance matrix [13], inter-channel time delay (ITD) and inter-channel level differences (ILD) [11], [14]. More recently, Chakrabarty et al. [16] proposed to use only the phase component of the spectrogram, avoiding explicit feature extraction. In the proposed method, we use both the magnitude and the phase component. Contrary to [16], which employed omnidirectional sensors only, general arrays with directional microphones additionally encode the DOA information in magnitude differences, while Ambisonics format especially encode directional information mainly in the magnitude component. f) All previous methods were evaluated on speech recordings that were synthetically spatialized and spatially static. We continue to use the static sound sources in the present work and extend them to a larger variety of sound events, such as impulsive and transient sounds.

II. Method

The block diagram of the proposed DOAnet is presented in Figure 1. The DOAnet takes multichannel audio as the input and first extracts the spectrograms of all the channels. The phases and the magnitudes of the spectrograms are mapped using a CRNN to two outputs sequentially. The first output, spatial pseudo-spectrum (SPS) is generated as a regression task, followed by the DOA estimates as a classification task. The DOA is defined by the azimuth ϕ and elevation λ with respect to the microphone and the SPS is the intensity of sound along the DOA given by $S(\phi, \lambda)$.

In this paper, we use discrete ϕ and λ by uniformly sampling the 2-D polar coordinate space, with a resolution of 10 degrees in both azimuth and elevation, resulting in 614 sampled directions. The SPS is computed at each sampled direction, whereas, a subset of 432 directions is used for DOA, where the elevations are limited between -60 and 60 degrees.

A. Feature extraction

The spectrogram is calculated for each of the audio channels whose sampling frequencies are 44100 Hz. A 2048-point discrete Fourier transform (DFT) is calculated on Hamming windows of 40 ms with 50 % overlap. We keep 1024 values of the DFT corresponding to the positive frequencies, without the zeroth bin. L frames of features, each containing 1024 magnitude and phase values of the DFT extracted in all the C channels, are stacked in a $L \times 1024 \times 2C$ 3-D tensor and used as the input to the proposed neural network. The $2C$ dimension results from ordering the magnitude component of all channels first, followed by the phase. We use a sequence length L of 100 (= 2 s) in this work.

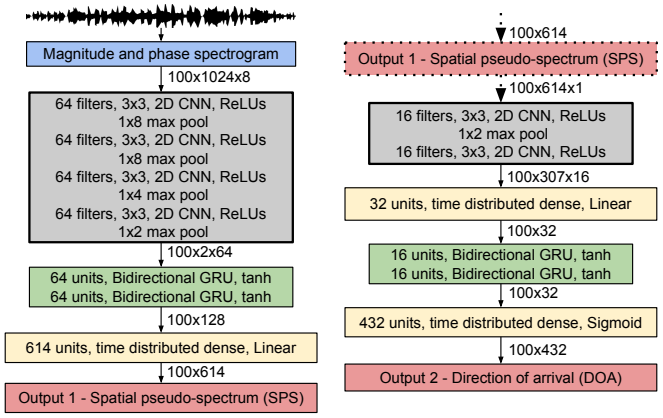


Fig. 1. DOAnet - neural network architecture for direction of arrival estimation of multiple sound sources.

B. Direction of arrival estimation network (DOAnet)

Local shift-invariant features are extracted from the input spectrogram tensor ($L \times 1024 \times 2C$ dimension) using CNN layers. In every CNN layer, the intra-channel time-frequency features are processed using a receptive field of 3×3, rectified linear unit (ReLU) activation and pad zeros to the resulting activation map to keep the output dimension equal to input. Batch normalization and max-pooling operation along frequency axis are performed after every CNN layer to reduce the final dimension to $L \times 2 \times N_C$, where N_C is the number of CNN filters in the last CNN layer. The CNN activations are reshaped to $L \times 2N_C$ keeping the time axis length unchanged and fed to RNN layers in order to learn temporal structure. Specifically, the bi-directional gated recurrent units (GRU) with tanh activation are used. Further, the RNN output is mapped to the first output, the SPS, in regression manner using FC layers with linear activation.

The SPS is further mapped to DOA estimates–the final output of the proposed method–using a similar CRNN network as above with two minor architectural changes. An FC layer is introduced between the CNN and RNN layers to reduce the dimension of the RNN output. Additionally, the output layer which predicts the DOA uses sigmoid activation in order to estimate more than one DOA for a given time frame. Each node in this output layer represents a direction in 2-D polar space. During testing, the probabilities at these nodes are thresholded with a value of 0.5, so that anything greater suggests the presence of a source in the direction or otherwise absence of source.

We refer to the combined architecture of SPS and DOA estimation in this work as DOAnet. The DOAnet is trained using the target SPS computed at each sampled direction, and for every time frame applying MUSIC (see Section III-B), and is represented using nonnegative real numbers. For the DOA output, the DOAnet aims to make a discrete decision about the presence of a source in a certain direction; and during training,

the DOAnet uses the ground truth DOAs utilized to synthesize the audio (see Section III-A).

The DOAnet was trained for 1000 epochs using Adam optimizer, mean squared error loss for SPS output and binary cross entropy loss for DOA output. The sum of the two losses was used for back propagation. Dropout was used after every layer and early stopping was used if the DOA metric (Section III-C) did not improve for 100 epochs. The DOAnet was implemented using Keras framework with Theano backend.

III. Evaluation

A. Dataset

In order to evaluate the proposed DOAnet, there are no publicly available real or synthetic datasets which consist of general sound events each associated with a 2D spatial coordinate. Since DNN-based methods need sufficiently large datasets to train on, most DNN-based methods proposed [11], [12], [14], [15], [16] have studied the performance on synthetic datasets. In similar fashion, we evaluate the proposed DOAnet on synthetic datasets about the same size as in the previous works.

We synthesize datasets consisting of static point sources associated with a spatial coordinate in the space in two contexts - anechoic and reverberant. For each context, three datasets are generated with no temporally overlapping sources ($O1$), maximum two overlapping sources ($O2$), and maximum three overlapping sound sources ($O3$). We refer to the anechoic context dataset as OxA and reverberant as OxR, where x denotes the number of overlapping sources. Each of these datasets has three cross-validation (CV) splits with 240 recordings for training and 60 for testing. Recordings are sampled at 44.1 kHz and 30 s long.

In order to generate these datasets, we use the isolated real-life sound event recordings from the DCASE 2016 task 2 [22]. This dataset consists of 11 sound event classes, each with 20 examples. The classes in this dataset included speech, coughing, door slam, page-turning, phone ringing and keyboard sounds. During CV, for each of the splits, we randomly chose disjoint sets of 16 and 4 examples for training and testing, amounting to 176 examples for training and 44 for testing. In order to synthesize a recording, a random subset of the 176 or 44 sound examples was chosen from the respective split. The subset size varied for each recording based on the chosen sound examples. We start synthesizing a recording by randomly choosing the beginning time of the first randomly chosen sound example within the first second of the recording. The next randomly chosen sound example is placed 250-500 ms after the end of the first sound example. On reaching the maximum recording length of 30 s, the process is repeated as many times as the number of required overlapping sound events.

Each of the sound examples were assigned a DOA randomly using the following conditions. All sound events were placed in a spatial grid of ten degrees resolution along both azimuth and elevation. Two temporally overlapping sound events have at least ten degrees of spatial separation to avoid spatial overlapping. The elevation was constrained within the range of [-60, 60] degrees, as most natural sound events occur in this range. Finally, for the anechoic dataset, the sound sources were randomly placed at a distance d in the range 1-10 m. For the reverberant dataset, the sound events were randomly placed inside a room of dimensions $10 \times 8 \times 4$ m with the microphone in the center of the room.

Spatialization for the anechoic case was done as following. Each point source signal s_i with DOA (ϕ_i, λ_i), was converted to Ambisonics format by multiplying the signal with the vector $\mathbf{y}(\phi_i, \lambda_i) = [Y_{00}(\phi_i, \lambda_i), Y_{1(-1)}(\phi_i, \lambda_i), Y_{10}(\phi_i, \lambda_i), Y_{11}(\phi_i, \lambda_i)]^{\mathrm{T}}$ of real orthonormalized spherical harmonics $Y_{nm}(\phi, \lambda)$. The complete anechoic sound scene multichannel recording $\mathbf{x}_{\mathrm{A}}$ was generated as $\mathbf{x}_{\mathrm{A}} = \sum_i g_i s_i \mathbf{y}(\phi_i, \lambda_i)$, with the gains $g_i < 1$ modeling the distance attenuation. Each entry of $\mathbf{x}_{\mathrm{A}}$ corresponds to one channel and $g_i = \sqrt{1/10^{d/d_{max}}}$, where $d_{max} = 10$ m is the maximum distance.

In the reverberant case, a fast geometrical acoustics simulator was used to model natural reverberation based on the rectangular room image-source model [23]. For each point source s_i with DOA in the dataset, K image sources were generated modeling reflections up to a predefined time-limit. Based on the room and its propagation properties, each image source was associated with a propagation filter h_{ik} and DOA (ϕ_k, λ_k) resulting in the spatial impulse response $\mathbf{h}_i = \sum_{k=1}^{K} h_{ik} \mathbf{y}(\phi_k, \lambda_k)$. The reverberant scene signal was finally generated by $\mathbf{x}_{\mathrm{R}} = \sum_i s_i * \mathbf{h}_i$, where $(*)$ denotes convolution of the source signal with the spatial impulse responses. The room absorption properties were adjusted to match reverberation times of typical office spaces. Three sets of testing data were generated with similar room size as training data (Room 1), 80% of room size ($8 \times 8 \times 4$ m) and reverberation time (Room 2), and 60% of room size ($8 \times 6 \times 4$ m) and reverberation time (Room 3).

B. Baseline

The proposed method to our knowledge is the first DNN-based implementation for 2D DOA estimation of multiple overlapping sound events. Thus in order to evaluate the complete features of the proposed DOAnet, we compare the performance with the conventional, high-resolution DOA estimator based on MUSIC. Similar to the SPS and DOA outputs estimated by the DOAnet, the MUSIC method also estimates SPS and DOA, thus allowing a direct one-to-one comparison.

The MUSIC SPS is based on a measure of orthogonality between the signal subspace (dominated by the source signals) of the spatial covariance matrix $\mathbf{C_s}$ and the noise subspace (dominated by diffuse and ambient sounds, late reverberation, and microphone noise). The spatial covariance matrix is calculated as $\mathbf{C_s} = \mathbb{E}_{f,t}\left[\mathbf{X}(f,t)\mathbf{X}(f,t)^{\mathrm{H}}\right]$, where spectrogram $\mathbf{X}(f,t)$ is a frequency f and time t dependent C-dimensional vector, where C is the number of channels, $^{\mathrm{H}}$ is the conjugate transpose and $\mathbb{E}_{f,t}$ denotes the expectation over f and t. For a sound scene with O number of sources, the MUSIC SPS

S_{GT} is obtained from $\mathbf{C_s}$ by first performing an eigenvalue decomposition on $\mathbf{C_s} = \mathbf{E\Lambda E^H}$. The sorted eigenvectors $\mathbf{E}$ (according to eigenvalues with decreasing magnitude) are further partitioned into the two aforementioned subspaces $\mathbf{E} = [\mathbf{U}_\mathrm{s}\ \mathbf{U}_n]$, where $\mathbf{U}_\mathrm{s}$ denotes the signal subspace and will be composed of O eigenvectors corresponding to the higher eigenvalues and the rest will form the noise subspace $\mathbf{U}_\mathrm{n}$. The S_{GT} along the direction (ϕ_i, λ_i) is now given by $S_{GT}(\phi_i, \lambda_i) = 1/(\mathbf{y}^\mathrm{T}(\phi_i, \lambda_i)\mathbf{U}_n\mathbf{U}_n^\mathrm{H}\mathbf{y}(\phi_i, \lambda_i))$. Finally, the source DOAs are found by selecting the directions (ϕ_i, λ_i) corresponding to the O largest peaks from S_{GT}.

C. Metric

The DOAnet estimated SPS ($S_E(\phi, \lambda)$) is evaluated with respect to the baseline MUSIC estimated ground truth ($S_{GT}(\phi, \lambda)$) using the SNR metric calculated as $SNR = 10\log_{10}(\sum_\phi \sum_\lambda S_{GT}(\phi, \lambda)^2 / \sum_\phi \sum_\lambda (S_E(\phi, \lambda) - S_{GT}(\phi, \lambda))^2)$.

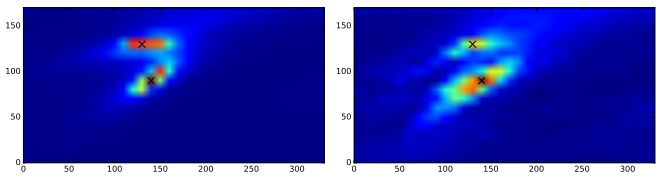
As the DOA metric we use the angle between the estimate DOA (defined by azimuth ϕ_E and elevation λ_E) and the ground truth DOA (ϕ_{GT}, λ_{GT}) used to synthesize the dataset in degrees. This is calculated as $\sigma = \arccos(\sin\phi_E \sin\phi_{GT} + \cos\phi_E \cos\phi_{GT} \cos(\lambda_{GT} - \lambda_E)) \cdot 180.0/\pi$. Further, to accommodate the scenario of unequal number of estimated and ground truth DOAs we calculate and report the minimum distance between them using the Hungarian algorithm [24] along with the percentage of frames in which the number of DOAs estimated were correct. The final metric for the entire dataset, referred as DOA error, is calculated by normalizing the minimum distance with the total number of estimated DOA's.

D. Evaluation procedure

The parameter tuning for DOAnet was performed on the $O1A$ test data, and the best configuration is as shown in Figure 1. This configuration has 677 K weights, and the same configuration is used in all of the following studies.

At test time, the SNR metric for SPS output of the DOAnet (S_E) is calculated with respect to SPS of baseline MUSIC (S_{GT}). The DOA metric for the DOAs predicted by DOAnet and baseline MUSIC are calculated with respect to the ground truth DOA used to synthesize the dataset.

In the above experiment, the baseline MUSIC algorithm uses the knowledge of the number of active sources. In order to have a fair evaluation, we test the DOAnet in a similar scenario where the number of sources is known. We use this knowledge to choose the top probabilities in prediction layer of the DOAnet instead of thresholding it with a value of 0.5.



(a) MUSIC estimated (b) DOAnet estimated

Fig. 2. SPS for two closely located sound sources. The black-cross markers represent the ground truth DOA. The horizontal axis is azimuth and vertical axis is elevation angle (in degrees)

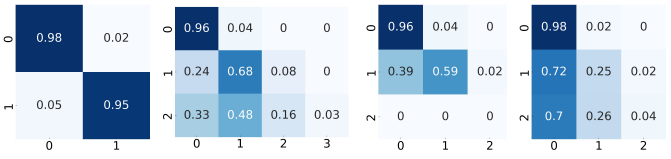
TABLE I
EVALUATION METRIC SCORES FOR THE SPATIAL POWER MAP AND DOAS ESTIMATED BY THE DOANET FOR DIFFERENT DATASETS.

	Anechoic			Reverberant (Room 1)		
Max. no. of overlapping sources	1	2	3	1	2	3
SPS SNR (in dB)	9.90	3.35	-0.26	3.11	1.24	0.13
DOA error with unknown number of active sources (threshold of 0.5)						
DOAnet	0.57	8.03	18.34	6.31	11.46	38.41
Correctly predicted frames (in %)	95.4	42.7	1.8	59.3	15.8	1.2
DOA error with known number of active sources						
DOAnet	1.14	27.52	49.30	12.61	38.98	67.07
MUSIC	2.29	8.60	28.66	25.80	57.33	91.72

IV. Results and Discussion

The results of the evaluations are presented in Table I. The high SNRs for SPS in both the contexts, with up to one and two overlapping sound events show that the SPS generated by DOAnet (S_E) is comparable with the baseline MUSIC SPS (S_{GT}). Figure 2 shows the S_E and the respective S_{GT} when two active sources are closely located. In the case of up to three overlapping sound events, the baseline MUSIC is already at its theoretical limit of estimating $N-1$ sources from N-dimensional signal space [25]. In practice, for $N-1$ sources only one noise subspace vector $\mathbf{U}_n$ is used to generate SPS, which for real signals is too weak for stable estimation. In the present evaluation of DOAnet which is trained with four-channel audio features and MUSIC SPS, for the case of three overlapping sound sources the SPS used is an unstable estimate resulting in poor training and consequently the results. With more than four-channels input, which the proposed DOAnet can easily extend to, it can potentially localize more than two sound sources simultaneously.

The DOA error for the proposed DOAnet when the number of active sources are unknown is presented in Table I. The DOAnet error is considerably better in comparison to the baseline MUSIC that uses the active sources knowledge for all datasets. However, the number of frames in which DOAnet produced the correct number of active sources were few. For example, in the case of anechoic recordings with up to two overlapping sound events, only 42.7% of the estimated frames had the correct number of DOA predictions. This prediction drops even drastically when the number of sources is three, due to the theoretical limit of MUSIC as explained previously, and consequently for the DOAnet as MUSIC SPS is used for training. Finally, the confusion matrix for the number of DOA estimates per frame for $O1$ and $O2$ datasets are visualized



(a) $O1A$ (b) $O2A$ (c) $O1R$ (d) $O2R$

Fig. 3. Confusion matrix for the number of DOA estimated per frame by the DOAnet. The horizontal axis is the DOAnet estimate, and the vertical axis is the ground truth.

TABLE II
EVALUATION SCORES FOR UNMATCHED REVERBERANT ROOM.

	Room 2		Room 3	
Max. no. of overlapping sources	1	2	1	2
SPS SNR (in dB)	3.53	1.49	3.49	1.46
DOAnet error (Unknown number of sources)				
DOAnet	3.44	6.88	4.59	10.89
Correctly predicted frames (in %)	46.2	14.3	49.7	14.1
DOA error (Known number of sources)				
DOAnet	8.60	32.10	9.17	33.82
MUSIC	31.52	58.47	33.25	60.76

in Figure 3. We skipped the confusion matrices for the $O3$ datasets as they were not meaningful for similar reasons as explained above.

With the knowledge of the number of active sources (Table I), the DOAnet performs considerably better than baseline MUSIC for all datasets other than the $O2A$ and $O3A$. The MUSIC DOA's were chosen using a 2D peak finder on the MUSIC SPS, whereas the DOA's in DOAnet were chosen by simply picking the top probabilities in the final DOA prediction layer. A smarter peak picking method from the DOAnet, or using the number of sources as an additional input can potentially result in better scores across all datasets. Further, the DOAnet error on unmatched reverberant data is presented in Table II. The performance of DOAnet is seen to be consistent in comparison to the matched reverberant data in Table I, and significantly better than the performance of MUSIC.

In this paper, since the baseline was chosen to be MUSIC, for a fair comparison the DOAnet was also trained using MUSIC SPS. In an ideal scenario, considering the DOAnet is trained using datasets for which the ground truth DOAs are known, we can generate accurate high-resolution SPS from the ground truth DOA's as per the required application and use them for training. Alternatively, the DOAnet can be trained without the SPS to directly generate the DOAs, it was only used in this paper to present the complete potential of the method in the limited paper space. In general, the above results show that the proposed DOAnet has the potential to learn the 2D direction information of multiple overlapping sound sources directly from the spectrogram of the input audio without the knowledge of the number of active sound sources. An exhaustive study with more detailed experiments including both synthetic and real datasets are planned for future work.

V. CONCLUSION

A convolutional recurrent neural network (DOAnet) was proposed for multiple source localization. The DOAnet was shown to learn the number of active sources directly from the input spectrogram, and estimate precise DOA in 2-D polar space. The method was evaluated on anechoic, matched and unmatched reverberant dataset. The proposed DOAnet performed considerably better than baseline MUSIC in most scenarios. Thereby showing the potential of DOAnet in learning highly computational algorithm without prior knowledge of the number of sources.

REFERENCES

[1] M. Woelfel and J. McDonough, "Distant speech recognition," in *Wiley*, 2009.

[2] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, 2014.

[3] A. Politis *et al.*, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, 2015.

[4] Y. Huang *et al.*, "Real-time passive source localization: a practical linear-correction least-squares approach," in *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, 2001.

[5] M. S. Brandstein and H. F. Silverman, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

[6] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, 1986.

[7] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 37, no. 7, 1989.

[8] J. H. DiBiase *et al.*, "Robust localization in reverberant rooms," in *Microphone Arrays*, 2001, pp. 157–180.

[9] A. O'Donovan *et al.*, "Imaging concert hall acoustics using visual and audio cameras," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[10] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, 2012.

[11] R. Roden *et al.*, "On sound source localization of speech signals using deep neural networks," in *Deutsche Jahrestagung für Akustik (DAGA)*, 2015.

[12] X. Xiao *et al.*, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[13] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016.

[14] A. Zermini *et al.*, "Deep neural network based audio source separation," in *International Conference on Mathematics in Signal Processing*, 2016.

[15] F. Vesperini *et al.*, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[16] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[17] H. Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*. Springer, 2007, vol. 348.

[18] T. N. Sainath *et al.*, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[19] M. Malik *et al.*, "Stacked convolutional and recurrent neural networks for music emotion recognition," in *Sound and Music Computing Conference (SMC)*, 2017.

[20] T. Sainath *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2017.

[21] S. Adavanne *et al.*, "Sound event detection using spatial features and convolutional recurrent neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[22] E. Benetos *et al.*, "Sound event detection in synthetic audio," http://www.cs.tut.fi/sgn/arg/dcase2016/, 2016.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," in *The Journal of the Acoustical Society of America*, vol. 65, no. 4, 1979.

[24] H. W. Kuhn, "The hungarian method for the assignment problem," in *Naval Research Logistics Quarterly*, no. 2, 1955, p. 8397.

[25] B. Ottersten *et al.*, "Exact and large sample maximum likelihood techniques for parameter estimation and detection in array processing," in *Radar Array Processing. Springer Series in Information Sciences*, 1993.

Publication V

Sharath Adavanne, Archontis Politis, Joonas Nikunen, Tuomas Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Network,"*IEEE Journal of Selected Topics in Signal Processing (JSTSP).* Volume 13, Issue 1, pp. 34-48, March 2019.

Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks

Sharath Adavanne, *Member, IEEE,* Archontis Politis, *Member, IEEE,* Joonas Nikunen, *Member, IEEE,* and Tuomas Virtanen, *Member, IEEE*

***Abstract*—In this paper, we propose a convolutional recurrent neural network for joint sound event localization and detection (SELD) of multiple overlapping sound events in three-dimensional (3D) space. The proposed network takes a sequence of consecutive spectrogram time-frames as input and maps it to two outputs in parallel. As the first output, the sound event detection (SED) is performed as a multi-label classification task on each time-frame producing temporal activity for all the sound event classes. As the second output, localization is performed by estimating the 3D Cartesian coordinates of the direction-of-arrival (DOA) for each sound event class using multi-output regression. The proposed method is able to associate multiple DOAs with respective sound event labels and further track this association with respect to time. The proposed method uses separately the phase and magnitude component of the spectrogram calculated on each audio channel as the feature, thereby avoiding any method- and array-specific feature extraction. The method is evaluated on five Ambisonic and two circular array format datasets with different overlapping sound events in anechoic, reverberant and real-life scenarios. The proposed method is compared with two SED, three DOA estimation, and one SELD baselines. The results show that the proposed method is generic and applicable to any array structures, robust to unseen DOA values, reverberation, and low SNR scenarios. The proposed method achieved a consistently higher recall of the estimated number of DOAs across datasets in comparison to the best baseline. Additionally, this recall was observed to be significantly better than the best baseline method for a higher number of overlapping sound events.**

***Index Terms*—Sound event detection, direction of arrival estimation, convolutional recurrent neural network**

I. Introduction

SOUND event localization and detection (SELD) is the combined task of identifying the temporal activities of each sound event, estimating their respective spatial location trajectories when active, and further associating textual labels with the sound events. Such a method can for example automatically describe social and human activities and assist the hearing impaired to visualize sounds. Robots can employ this for navigation and natural interaction with surroundings [1–4]. Smart cities, smart homes, and industries could use it for audio surveillance [5–8]. Smart meeting rooms can recognize speech among other events and use this information to beamform and enhance the speech for teleconferencing or for robust automatic speech recognition [9–13]. Naturalists could use it for bio-diversity monitoring [14–16]. Further, in virtual reality (VR) applications with 360° audio SELD can be used to assist the user in visualizing sound events.

A. Sound event detection

The SELD task can be broadly divided into two sub-tasks, sound event detection (SED) and sound source localization. SED aims at detecting temporally the onsets and offsets of sound events and further associating textual labels to the detected events. The sound events in real-life most often overlap with other sound events in time and the task of recognizing all the overlapping sound events is referred as polyphonic SED. The SED task in literature has most often been approached using different supervised classification methods that predict the framewise activity of each sound event class. Some of the classifiers include Gaussian mixture model (GMM) - hidden Markov model (HMM) [27], fully connected (FC) neural networks [28], recurrent neural networks (RNN) [29–32], and convolutional neural networks (CNN) [33, 34]. More recently state-of-the-art results were obtained by stacking CNN, RNN and FC layers consecutively, referred jointly as the convolutional recurrent neural network (CRNN) [35–39].

Lately, in order to improve recognition of overlapping sound events, several multichannel SED methods have been proposed [39–43] and these were among the top performing methods in the real-life SED task of DCASE 2016[1] and 2017[2] evaluation challenges. More recently, we studied the SED performance on identical sound scenes captured using single, binaural and first-order Ambisonics (FOA) microphones [35], where the order denotes the spatial resolution of the format and the first order corresponds to four channels. The results showed

S. Adavanne, J. Nikunen and T.Virtanen are with the Signal Processing Laboratory, Tampere University of Technology, Finland, e-mail: firstname.lastname@tut.fi

A. Politis is with the Department of Signal Processing and Acoustics, Aalto University, Finland, e-mail: archontis.politis@aalto.fi

The research leading to these results has received funding from the European Research Council under the European Unions H2020 Framework Program through ERC Grant Agreement 637422 EVERYSOUND. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-sound-event-detection-in-real-life-audio#system-characteristics

[2] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-sound-event-detection-in-real-life-audio-results#system-characteristics

TABLE I
SUMMARY OF DNN BASED DOA ESTIMATION METHODS IN THE LITERATURE. THE AZIMUTH AND ELEVATION ANGLES ARE DENOTED AS 'AZI' AND 'ELE', DISTANCE AS 'DIST', 'X' AND 'Y' REPRESENT THE DISTANCE ALONG THE RESPECTIVE CARTESIAN AXIS. 'FULL' REPRESENTS THE ESTIMATION IN THE COMPLETE RANGE OF THE RESPECTIVE FORMAT, AND 'REGRESSION' REPRESENTS THE CLASSIFIER ESTIMATION TYPE.

Approach	Input feature	Output format	Sources	DNN	Array	SELD
Chakrabarty et al. [17, 18]	Phase spectrum	azi	1, multiple	CNN	Linear	×
Yalta et al. [3]	Spectral power	azi (Full)	1	CNN Resnet	Robot	×
Xiao et al. [19]	GCC	azi (Full)	1	FC	Circular	×
Takeda et al. [1, 2]	Eigen vectors of spatial covariance matrix	azi (Full)	1, 2	FC	Robot	×
He et al. [4]	GCC	azi (Full)	Multiple	CNN	Robot	×
Hirvonen [20]	Spectral power	azi (Full) for each class	Multiple	CNN	Circular	✓
Yiwere et al. [21]	ILD, cross-correlation	azi and dist	1	FC	Binaural	×
Ferguson et al. [22]	GCC, cepstrogram	azi and dist (regression)	1	CNN	Linear	×
Vesperini et al. [23]	GCC	x and y (regression)	1	FC	Distributed	×
Sun et al. [24]	GCC	azi and ele	1	PNN	Cartesian	×
Adavanne et al. [25]	Phase and magnitude spectrum	azi and ele (Full)	Multiple	CRNN	Generic	×
Roden et al. [26]	ILD, ITD, phase and magnitude spectrum	azi, ele and dist (separate NN)	1	FC	Binaural	×
Proposed	Phase and magnitude spectrum	azi and ele (Full, regression) for each class	Multiple	CRNN	Generic	✓

that the recognition of overlapping sound events improved with increase in spatial sampling, and the best performance was obtained with FOA.

B. Sound source localization

Sound source localization is the task of determining the direction or position of a sound source with respect to the microphone. In this paper, we only deal with the estimation of the sound event direction, generally referred as direction-of-arrival (DOA) estimation. The DOA methods in literature can be broadly categorized into parametric- and deep neural network (DNN)-based approaches. Some popular parametric methods are based on time-difference-of-arrival (TDOA) [44], the steered-response-power (SRP) [45], multiple signal classification (MUSIC) [46], and the estimation of signal parameters via rotational invariance technique (ESPRIT) [47]. These methods vary in terms of algorithmic complexity, constraints in array geometry, and model assumptions on the acoustic scenarios. Subspace methods like MUSIC can be applied with different array types and can produce high-resolution DOA estimates of multiple sources. On the other hand, subspace methods require a good estimate of the number of active sources that may be hard to obtain, and they have been found sensitive to reverberant and low signal-to-noise (SNR) scenarios [48].

Recently, DNN-based methods were employed to overcome some of the drawbacks of parametric methods, while being robust towards reverberation and low SNR scenarios. Additionally, implementing the localization task in the DNN framework allows seamless integration into broader DNN tasks such as SELD [20], robots can use it for sound source based navigation and natural interaction in multi-speaker scenarios [1–4]. A summary of the most recent DNN-based DOA estimation methods is presented in Table I. All these methods estimate DOAs for static point sources and were shown to perform equally or better than the parametric methods in reverberant scenarios. Further, methods [4, 18, 20, 25] proposed to simultaneously detect DOAs of overlapping sound events by estimating the number of active sources from the data itself. Most methods used a classification approach, thereby estimating the source presence likelihood at a fixed set of angles, while [22, 23] used a regression approach and let the DNN produce continuous output.

All of the past works were evaluated on different array geometries, making a direct performance comparison difficult. Most of the methods estimated full azimuth ('Full' in Table I) using microphones mounted on a robot, circular and distributed arrays, while the rest of the methods used linear arrays thereby estimating only the azimuth angles in a range of $180°$. Although few of the existing methods estimated the azimuth and elevation jointly [24, 25], most of them estimated only the azimuth angle [1–4, 17–20]. In particular, we studied the joint estimation of azimuth and elevation angles in [25], this was enabled by the use of Ambisonic signals (FOA) obtained using a spherical array. Ambisonics are also known as spherical harmonic (SH) signals in the array processing literature, and they can be obtained from various array configurations such as circular or planar (for 2D capture) and spherical or volumetric (for 3D capture) using an appropriate linear transform of the recordings [49]. The same ambisonic channels have the same spatial characteristics independent of the recording setup, and hence, studies on such hardware-independent formats make the evaluation and results more easily comparable in the future.

Most of the previously proposed DNN-based DOA estimation methods that relied on a single array or distributed arrays of omnidirectional microphones, captured source location information mostly in phase- or time-delay differences between the microphones. However, compact microphone arrays with full azimuth and elevation coverage, such as spherical microphone arrays, rely strongly on the directionality of the sensors to capture spatial information, this reflects mainly in the magnitude differences between channels. Motivated by this fact we proposed to use both the magnitude and phase component of the spectrogram as input features in [25]. Thus making the DOA estimation method [25] generic to array configuration by avoiding method-specific feature extractions like inter-aural

level difference (ILD), the inter-aural time difference (ITD), generalized cross-correlation (GCC) or eigenvectors of spatial covariance matrix used in previous methods (Table I).

C. Joint localization and detection

In the presence of multiple overlapping sound events, the DOA estimation task becomes the classical tracking problem of associating correctly the multiple DOA estimates to respective sources, without necessarily identifying the source [50, 51]. The problem is further extended for the polyphonic SELD task if the SED and DOA estimation are done separately, resulting in the data association problem between the recognized sound events and the estimated DOAs [13]. One solution to the data association problem is to jointly predict the SED and DOA. In this regard, to the best of the authors' knowledge, [20] is the only DNN-based method which performs SELD. Other works combining SED and parametric DOA estimation include [6, 13, 52, 53]. Lopatka et al. [53] used a 3D sound intensity acoustic vector sensor, with MPEG-7 spectral and temporal features along with a support vector machine classifier to estimate DOA along azimuth for five classes of non-overlapping sound events. Butko et al. [13] used distributed microphone arrays to recognize 14 different sound events with an overlap of two at a time, using a GMM-HMM classifier, and localized them inside a meeting room using the SRP method. Chakraborty et al. [52] replaced SRP-based localization in [13] with a sound-model-based localization, thereby fixing the data association problem faced in [13]. In contrast, Hirvonen [20], extracted the frame-wise spectral power from each microphone of a circular array and used a CNN classifier to map it to eight angles in full azimuth for each sound event class in the dataset. In this output format, the resolution of azimuth is limited to the trained directions and the performance of unseen DOA values is unknown. For larger datasets with a higher number of sound events and increased resolution along azimuth and elevation directions, this approach results in a large number of output nodes. Training such a DNN with a large number of output nodes where the number of positive class labels per frame is one or two with respect to a high number of negative class labels poses challenges of an imbalanced dataset. Additionally, training such a large number of classes requires a huge dataset with enough examples for each class. On the other hand, this output format allows the network to simultaneously recognize more than one instance of the same sound event in a given time frame, at different locations.

D. Contributions of this paper

In general, the number of existing SELD methods is limited [6, 13, 20, 52, 53], with only one published DNN-based approach [20]. On the other hand, there are several DNN-based methods in the literature for the SELD sub-tasks of SED and DOA estimation. Yet, there is no comprehensive work published that studies the various choices affecting the performance of these DNN-based SED, DOA and SELD methods, compare them with multiple competitive baselines, and evaluate them over a wide range of acoustic conditions. Besides, with respect to the SELD task, the existing methods [6, 13, 52, 53] localize up to one or maximum two overlapping sound events and do not scale to a higher number of overlapping sources. Further, the only DNN-based SELD method [20] localizes sound events exclusively at a predefined grid of directions and requires a large number of output classes for a higher number of sound event labels and increased spatial resolution. Additionally, all the above SELD approaches use method-specific features and hence not independent of input array structure.

In contrast to existing SELD methods, this paper presents novelty in two broad areas: the proposed SELD method, and the exhaustive evaluation studies presented. The novelty of the proposed SELD method is as follows. It is the first method that addresses the problem of localizing and recognizing more than two overlapping sound events simultaneously and tracking their activity with respect to time. The proposed method is able to localize sources at any azimuth and elevation angles while being robust to unseen spatial locations, reverberation, and ambiance. Further, the method itself is generic enough to learn to perform SELD from any input array structure. Specifically, as our method, we propose to use the polyphonic SED output [39] as a confidence measure for choosing the DOAs estimated in a regression manner. By this approach, we not only extend the state-of-the-art polyphonic SED performance [39] for polyphonic SELD but also tackle the data-association problem faced due to the polyphony in SELD tasks [13]. As the second broad area of novelty, we present the performance of the proposed method with respect to various design choices made such as the DNN architecture, input feature and DOA output format. Additionally, we also present the comprehensive results of the proposed method with respect to six baselines (two SED, three DOA estimation, and one SELD baseline) evaluated on seven datasets with different acoustic conditions (anechoic and reverberant scenarios with simulated and real-life impulse responses), array configurations (Ambisonic and circular array) and the number of overlapping sound events.

In order to facilitate reproducibility of research, the proposed method and all the datasets used have been made publicly available[3]. Additionally, the real-life impulse responses used to simulate datasets have also been published to enable users to experiment with custom sound events.

The rest of the paper is organized as follows. In Section II, we describe the proposed SELD method and the training procedure. In Section III, we describe the datasets, the baseline methods, the metrics and the experiments carried out for evaluating the proposed method. The experimental results on the evaluation datasets are presented, compared with baselines and discussed in Section IV. Finally, we summarize the conclusions of the work in Section V.

II. Method

The block diagram of the proposed method for SELD is presented in Figure 1a. The input to the method is the multichannel audio. The phase and magnitude spectrograms

[3]https://github.com/sharathadavanne/seld-net

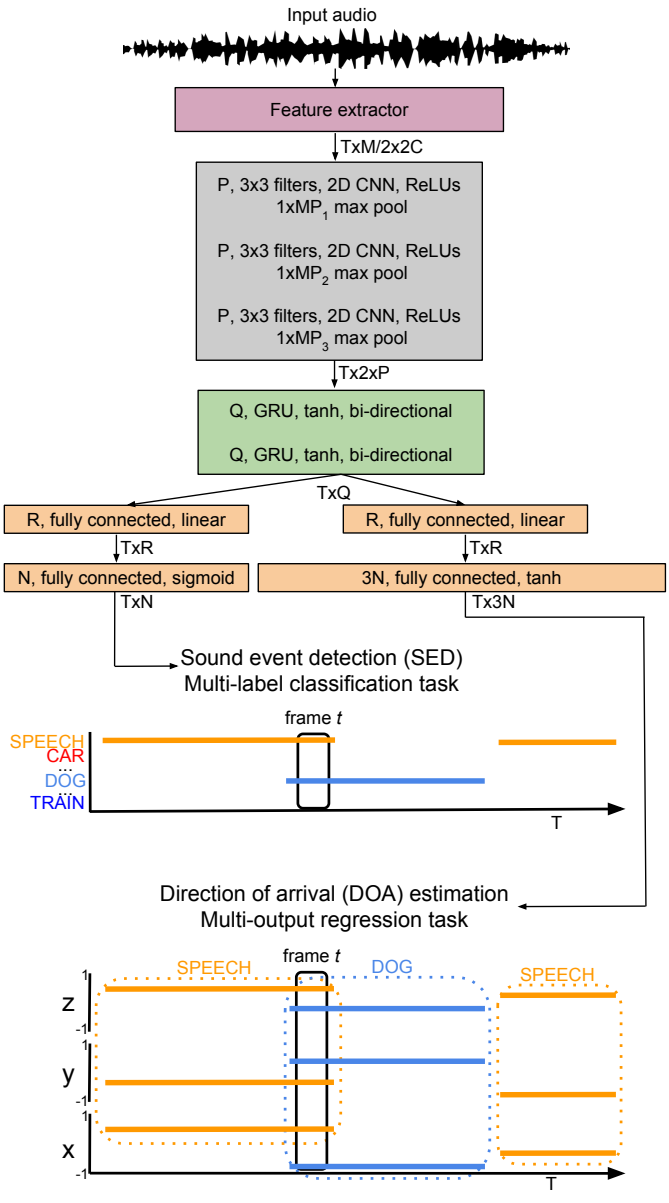
are extracted from each audio channel and are used as separate features. The proposed method takes a sequence of features in consecutive spectrogram frames as input and predicts all the sound event classes active for each of the input frames along with their respective spatial location, producing the temporal activity and DOA trajectory for each sound event class. In particular, a CRNN is used to map the feature sequence to the two outputs in parallel. At the first output, SED is performed as a multi-label classification task, allowing the network to simultaneously estimate the presence of multiple sound events for each frame. At the second output, DOA estimates in the continuous 3D space are obtained as a multi-output regression task, where each sound event class is associated with three regressors that estimate the 3D Cartesian coordinates x, y and z of the DOA on a unit sphere around the microphone. The SED output of the network is in the continuous range of [0 1] for each sound event in the dataset, and this value is thresholded to obtain a binary decision for the respective sound event activity as shown in Figure 1b. Finally, the respective DOA estimates for these active sound event classes provide their spatial locations. The detailed description of the feature extraction and the proposed method is explained in the following sections.

A. Feature extraction

The spectrogram is extracted from each of the C channels of the multichannel audio using an M-point discrete Fourier transform (DFT) on Hamming window of length M and 50% overlap. The phase and magnitude of the spectrogram are then extracted and used as separate features. Only the $M/2$ positive frequencies without the zeroth bin are used. The output of the feature extraction block in Figure 1a is a feature sequence of T frames, with an overall dimension of $T \times M/2 \times 2C$, where the $2C$ dimension consists of C magnitude and C phase components.

B. Neural network architecture

The output of the feature extraction block is fed to the neural network as shown in Figure 1a. In the proposed architecture the local shift-invariant features in the spectrogram are learned using multiple layers of 2D CNN. Each CNN layer has P filters of $3 \times 3 \times 2C$ (as in [25]) dimensional receptive fields acting along the time-frequency-channel axis with a rectified linear unit (ReLU) activation. The use of filter kernels spanning all the channels allows the CNN to learn relevant inter-channel features required for localization, whereas the time and frequency dimensions of the kernel allows learning relevant intra-channel features suitable for both the DOA and SED tasks. After each layer of CNN, the output activations are normalized using batch normalization [54], and the dimensionality is reduced using max-pooling (MP_i) along the frequency axis, thereby keeping the sequence length T unchanged. The output after the final CNN layer with P filters is of dimension $T \times 2 \times P$, where the reduced frequency dimension of 2 is a result of max-pooling across CNN layers (see Section IV-1).



(a) SELDnet

Sound event class	SED output	Sound event activity	DOA estimates x	y	z	
SPEECH	0.8		0.4	-0.4	0.5	Sound event active
CAR	0.1	Threshold > 0.5	0.3	-0.1	0.0	
...	0.2		0.1	0.2	0.1	
DOG	0.7		-0.8	0.4	-0.2	Sound event inactive
...	...		...	...	...	
TRAIN	0.1		0.1	0.0	-0.1	

(b) SELDnet output

Fig. 1. a) The proposed SELDnet and b) the frame-wise output for frame t in Figure a). A sound event is said to be localized and detected when the confidence of the SED output exceeds the threshold.

The output activation from CNN is further reshaped to a T frame sequence of length $2P$ feature vectors and fed to bidirectional RNN layers which are used to learn the temporal context information from the CNN output activations. Specifically, Q nodes of gated recurrent units (GRU) are used in each layer with tanh activations. This is followed by two branches of FC layers in parallel, one each for SED and DOA estimation. The FC layers share weights across time steps. The first FC layer in both the branches contains R nodes each with linear activation. The last FC layer in the SED branch consists of N nodes with sigmoid activation, each corresponding to one of the N sound event classes to be detected. The use of sigmoid activation enables multiple classes to be active simultaneously. The last FC layer in the DOA branch consists of $3N$ nodes with tanh activation, where each of the N sound event classes is represented by 3 nodes corresponding to the sound event location in x, y, and z, respectively. For a DOA estimate on a unit sphere centered at the origin, the range of location along each axes is $[-1, 1]$, thus we use the tanh activation for these regressors to keep the output of the network in a similar range.

We refer to the above architecture as SELDnet. The SED output of the SELDnet is in the continuous range of $[0, 1]$ for each class, while the DOA output is in the continuous range of $[-1, 1]$ for each axes of the sound class location. A sound event is said to be active, and its respective DOA estimate is chosen if the SED output exceeds the threshold of 0.5 as shown in Figure 1b. The network hyperparameters are optimized based on cross-validation as explained in Section III-D1.

C. Training procedure

In each frame, the target values for each of the active sound events in the SED branch output are one while the inactive events are zero. Similarly, for the DOA branch, the reference DOA x, y, and z values are used as targets for the active sound events and $x = 0$, $y = 0$, and $z = 0$ is used for inactive events. A binary cross-entropy loss is used between the SED predictions of SELDnet and reference sound class activities, while a mean square error (MSE) loss is used for the DOA estimates of the SELDnet and the reference DOA. By using the MSE loss for DOA estimation in 3D Cartesian coordinates we truly represent the distance between two points in space. The distance between two points (x_1, y_1, z_1) and (x_2, y_2, z_2) in 3D space is given by $\sqrt{SE}$, where $SE = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2$, while the MSE between the same points is given by $SE/3$. Thus the MSE loss is simply a scaled version of the distance in 3D space, and reducing the MSE loss implies the reduction in the distance between the two points.

Theoretically, the advantage of using Cartesian coordinates instead of azimuth and elevation for regression can be observed when predicting DOA in full azimuth and/or full elevation. The angles are discontinuous at the wrap-around boundary (for example the $-180°$, $180°$ boundary for azimuth), while the Cartesian coordinates are continuous. This continuity allows the network to learn better. Further experiments on this are discussed in Section III-D.

We train the SELDnet with a weighted combination of MSE and binary cross-entropy loss for 1000 epochs using Adam optimizer with default parameters as used in the original paper [55]. Early stopping is used to control the network from over-fitting to training split. The training is stopped if the SELD score (Section III-C) on the test split does not improve for 100 epochs. The network was implemented using Keras library [56] with TensorFlow [57] backend.

III. Evaluation

A. Datasets

The proposed SELDnet is evaluated on seven datasets that are summarized in Table II. Four of the datasets are synthesized with artificial impulse responses (IR), that consists of anechoic and reverberant scenarios virtually recorded both with a circular array and in the Ambisonics format. Three of the datasets are synthesized with real-life impulse responses, recorded with a spherical array and encoded into the Ambisonics format. All the datasets consist of stationary point sources each associated with a spatial coordinate. The synthesis procedure in all the datasets consists of mixing isolated sound event instances at different spatial locations, since this allows producing the reference event locations and times of activity for evaluation and training of the methods.

1) TUT Sound Events 2018 - Ambisonic, Anechoic and Synthetic Impulse Response (ANSYN) dataset: This dataset consists of spatially located sound events in an anechoic environment synthesized using artificial IRs. It comprises three subsets: no temporally overlapping sources ($O1$), maximum two temporally overlapping sources ($O2$) and maximum three temporally overlapping sources ($O3$). Each of the subsets consists of three cross-validation splits with 240 training and 60 testing FOA format recordings of length 30 s sampled at 44100 Hz. The dataset is generated using the 11 isolated sound event classes from the DCASE 2016 task 2 dataset [58] such as speech, coughing, door slam, page-turning, phone ringing and keyboard. Each of these sound classes has 20 examples, of which 16 are randomly chosen for the training set and the rest four for the testing set, amounting to 176 examples from 11 classes for training, and 44 for testing. During synthesis of a recording, a random collection of examples are chosen from the respective set and are randomly placed in a spatial grid of $10°$ resolution along azimuth and elevation, such that two overlapping sound events are separated by $10°$, and the elevation is in the range of $[-60°, 60°)$. In order to have a variability of amplitude, the sound events are randomly placed at a distance ranging from 1 to 10 m with 0.5 m resolution from the microphone. More details regarding the synthesis can be found in [25].

2) TUT Sound Events 2018 - Ambisonic, Reverberant and Synthetic Impulse Response (RESYN) dataset: This dataset is synthesized with the same details as the ambisonic ANSYN dataset, with the only difference being that the sound events are spatially placed within a room using the image source method [59]. Specifically, the microphone is placed at the center of the room, and the sound events are randomly placed around the microphone, with their distance ranging from 1 m from the microphone to the respective end of the room at 0.5 m resolution. The three cross-validation splits

TABLE II
SUMMARY OF DATASETS

Audio format	Sound scene	Impulse response	Dataset acronym	Train/Test, notes
Ambisonic (four channel)	Anechoic	Synthetic	ANSYN	240/60
	Reverberant		RESYN	
		Real life	REAL	
			REALBIG	600/150
			REALBIGAMB	600/150, ambiance
Circular array (eight channel)	Anechoic	Synthetic	CANSYN	240/60
	Reverberant		CRESYN	

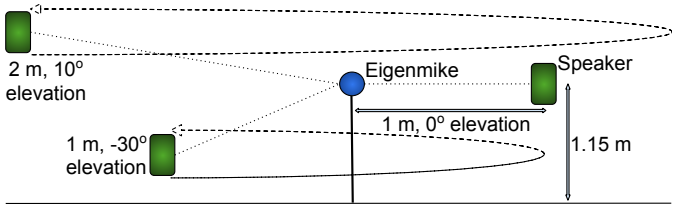


Fig. 2. Recording real-life impulse responses for sound scene generation. A person walks around the Eigenmike[4] holding a Genelec loudspeaker[5] playing a maximum length sequence at different elevation angles and distances.

of each of the three subsets $O1$, $O2$ and $O3$ are generated for a moderately reverberant room of size $10 \times 8 \times 4$ m (Room 1), with reverberation times 1.0, 0.8, 0.7, 0.6, 0.5, and 0.4 s per each octave band, and 125 Hz–4 kHz band center frequencies. Additionally, to study the performance in mismatched reverberant scenarios, testing splits are generated for two different sized rooms: room 2 that is 80% the volume ($8 \times 8 \times 4$ m) and reverberation-time per band of room 1, and room 3 that is 125% the volume ($10 \times 10 \times 4$ m) and reverberation-time per band of room 1. In order to remove any ambiguity while comparing the performance difference of room 1 with room 2 and 3, we keep the sound events and their respective spatial locations in room 2 and 3 identical to the testing split of room 1. But the individual sound events whose distance from the microphone exceeded the room size were reassigned a new distance within the room. Further details on the reverberant synthesis can be read in [25].

3) TUT Sound Events 2018 - Ambisonic, Reverberant and Real-life Impulse Response (REAL) dataset: In order to study the performance of SELDnet in a real-life scenario, we generated a dataset by collecting impulse responses from a real environment using the Eigenmike[4] spherical microphone array. For the IR acquisition, we used a continuous measurement signal as in [60]. The measurement was done by slowly moving a Genelec G Two loudspeaker[5] continuously playing a maximum length sequence around the array in circular trajectory in one elevation at a time, as shown in Figure 2. The playback volume was set to be 30 dB greater than the ambient sound level. The recording was done in a corridor inside the university with classrooms around it.

The moving-source IRs were obtained by a freely available tool from CHiME challenge [61] which estimates the time-varying responses in STFT domain by forming a least-squares regression between the known measurement signal and the far-field recording independently at each frequency. The IR for any azimuth within one trajectory can be analyzed by assuming block-wise stationarity of acoustic channel. The average angular speed of the loudspeaker in the measurements was 6°/s and we used a block size of 860 ms (81 STFT frames with analysis frame size of 1024 with 50 % overlap and sample rate $F_s = 48$ kHz) for estimation of IR of length 170 ms (16 STFT frames).

The IRs were collected at elevations $-40°$ to $40°$ with $10°$ increments at 1 m from the Eigenmike and at elevations $-20°$ to $20°$ with $10°$ increments at 2 m. For the dataset creation, we analyzed the DOA of each time frame using MUSIC and extracted IRs for azimuthal angles at $10°$ resolution (36 IRs for each elevation). The IR estimation tool [61] was applied independently on all 32 channels of the Eigenmike.

In order to synthesize the sound scene from the estimated IRs, we used isolated real-life sound events from the urban-sound8k dataset [62]. This dataset consists of 10 sound event classes such as: air conditioner, car horn, children playing, dog barking, drilling, engine idling, gunshot, jackhammer, siren and street music. Among these, we did not include children playing and air conditioner classes since these can also occur in our ambiance recording which we use as background recording in dataset REALBIGAMB (Section III-A5). From the sound examples in urbansound8k, we only used the ones marked as foreground in order to have clean isolated sound events. Similarly to the other datasets used in this paper, we used the splits 1, 8 and 9 provided in the urbansound8k as the three cross-validation splits. These splits were chosen as they had a good number of examples for all the chosen sound event classes after selecting only the foreground examples. The final selected examples varied in length from 100 ms to 4 s and amount to 15671.5 seconds from 4542 examples.

During the sound scene synthesis, we randomly chose a sound event example and associated it with a random distance among the collected ones, azimuth and elevation angle. The sound event example was then convolved with the respective IR for the given distance, azimuth and elevation to spatially position it. Finally, after positioning all the sound events in a recording we converted this multichannel audio to FOA format. The transform of the microphone signals to FOA was performed using the tools published in [63]. In total, we generated 300 such 30 s recordings in a similar fashion as ANSYN and RESYN with 240 of them earmarked for training and 60 for testing. Similar to the ANSYN recordings we also generated three subsets $O1$, $O2$ and $O3$ with a different number of overlapping sound events.

4) TUT Sound Events 2018 - Ambisonic, Reverberant and Real-life Impulse Response big (REALBIG) dataset: In order to study the performance of SELDnet with respect to the size of the dataset, we generated for each of three ambisonic REAL subsets a 750 recordings REALBIG subset of 30 s length, with 600 for training and 150 for testing.

5) TUT Sound Events 2018 - Ambisonic, Reverberant, Real-life Impulse Response and Ambiance big (REALBIGAMB) dataset: Additionally, to simulate a real sound-scene we

[4]https://mhacoustics.com/products
[5]https://www.genelec.com/home-speakers/g-series-active-speakers

recorded 30 min of ambient sound to use as background noise in the same location as the IR recordings without changing the setup. We mixed randomly chosen segments of the recorded ambiance at three different SNRs: 0, 10 and 20 dB for each of the three ambisonic REALBIG subsets and refer to it as REALBIGAMB subsets. The ambiance used for the testing set was kept separate from the training set.

6) TUT Sound Events 2018 - Circular array, Anechoic and Synthetic Impulse Response (CANSYN) dataset: To study the performance of SELDnet on generic array configurations, similarly to the SELD baseline method [20] (Section III-B3), we synthesized the ANSYN recordings for a circular array of radius 5 cm with eight omnidirectional microphones at $0, 45, 90, 135, 180, 225, 270, 315^o$, and the array plane parallel to the ground, and refer to it as CANSYN. It is an exact replica of the ANSYN dataset in terms of the synthesized sound events except for the microphone array setup, and hence the number of channels. Similar to ANSYN, the CANSYN dataset has three subsets with a different number of overlapping sound events each with three cross-validation splits.

7) TUT Sound Events 2018 - Circular array, Reverberant and Synthetic Impulse Response (CRESYN) dataset: Similar to the CANSYN dataset, we synthesize the circular array version of ambisonic RESYN room 1 dataset, referred as CRESYN. During synthesis, the circular microphone array is placed in the center of the room, and the array plane parallel to the floor.

B. Baseline methods

The SELDnet is compared with six different baselines as summarized in Table III: two SED baselines (single- and multichannel), three DOA baselines (parametric and DNN-based), and a SELD baseline.

1) SED baseline: The SED capabilities of the proposed SELDnet is compared with the existing state-of-the-art multi-channel SED method [39], referred here as MSEDnet. MSEDnet is easily scalable to any number of input audio channels and won [38] the recently concluded real-life SED task in DCASE 2017 [64]. In particular, it won the top two positions among 34 submissions, first using single-channel mode (referred as SEDnet) and a close second using multichannel mode. The SED performance of SELDnet is compared with both the single- and the multichannel modes of MSEDnet.

In the original MSEDnet implementation [39] the input is a sequence of log mel-band energy (40-bands) frames, that are mapped to an equal-length sequence of sound event activities. The SED metrics (Section III-C) for MSEDnet did not change much on using phase and magnitude components of the STFT spectrogram instead of log mel-band energies. Hence, in order to have a one-to-one comparison with SELDnet, we use the phase and magnitude components of the STFT spectrogram for MSEDnet in this paper. We train the MSEDnet for 500 epochs and use early stopping when SED score (Section III-C) stops improving for 100 epochs.

2) DOA baseline: The DOA estimation performance of the SELDnet is evaluated with respect to three baselines. As a parametric baseline, we use MUSIC [46] and as DNN-based baselines, we use the recently proposed DOAnet [25] that estimates DOAs in 3D and [18] that estimates only the DOA azimuth angle referred as AZInet.

TABLE III
BASELINE AND PROPOSED METHOD SUMMARY

Task	Acronym	Notes	Datasets evaluated
SED	SEDnet [39]	Single channel	All
	MSEDnet [39]	Multichannel	
DOA	MUSIC*	Azi and ele	All except CANSYN and CRESYN
	DOAnet [25]	Azi and ele	
	AZInet [18]	Azi	CANSYN and CRESYN
SELD	HIRnet [20]	Azi	
	SELDnet-azi	Azi	All
	SELDnet	Azi and ele	

*Parametric, all other methods are DNN based

i) MUSIC: is a versatile high-resolution subspace method that can detect multiple narrowband source DOAs and can be applied to generic array setups. It is based on a subspace decomposition of the spatial covariance matrix of the multichannel spectrogram. For a broadband estimation of DOAs, we combine narrowband spatial covariance matrices over three frames and frequency bins from 50 to 8000 Hz. The steering vector information required to produce the MUSIC pseudo-spectrum from which the DOAs are extracted is adapted to the recording system under use, meaning uniform circular array steering vectors for CANSYN and CRESYN datasets, and real SH vectors for all the other ambisonic datasets.

MUSIC requires a good estimate of the number of active sound sources in order to estimate their DOAs. In this paper, we use MUSIC with the number of active sources taken from the reference of the dataset. Hence, the DOA estimation results of MUSIC can be considered as the best possible for the given dataset and serve as a benchmark for DOA estimation with and without the knowledge of the number of active sources. For a detailed description on MUSIC and other subspace methods, the reader is referred to [65], while for application of MUSIC to SH signals similar to this work, please refer to [66].

ii) DOAnet: Among the recently proposed DNN-based DOA estimation methods listed in Table I, the only method that attempts DOA estimation of multiple overlapping sources in 3D space is the DOAnet [25]. Hence, DOAnet serves as a suitable baseline to compare against the DOA estimation performance of the proposed SELDnet. DOAnet is based on a similar CRNN architecture, the input to which is a sequence of multichannel phase and magnitude spectrum frames. It considers DOA estimation as a multi-label classification task by directional sampling with a resolution of 10° along azimuth and elevation and estimating the likelihood of a sound source being active in each of these points.

iii) AZInet: Among the DOA-only estimation methods listed in Table I, apart from the DOAnet [25], methods [18] and [4] are the only ones which attempt simultaneous DOA estimation of overlapping sources. Since [4] is evaluated on a dataset collected using microphones mounted on a humanoid robot, it is difficult to replicate the setup. Hence in this paper, we use the AZInet evaluated on a linear array in [18] as the baseline. The AZInet is a CNN-based method that uses the phase component of the spectrogram of each channel as input,

and maps it to azimuth angles in the range $0°$ to $180°$ at $5°$ resolution as a multi-label classification task. AZInet uses only the phase spectrogram since the dataset evaluated on employs omnidirectional microphones, which for compact arrays and sources in the far-field, preserve directional information in inter-channel phase differences. Thus, although the evaluation in [18] was carried out on a linear array, the method is generic to any omnidirectional array under these conditions. Further, in order to have a direct comparison, we extend the output of AZInet to full-azimuth with $10°$ resolution and reduce the output of SELDnet to generate only the azimuth, i.e., we only estimate x and y coordinates of the DOA (SELDnet-azi). To enable this full-azimuth estimation we use the circular array with omnidirectional microphones datasets CANSYN and CRESYN.

3) SELD baseline (HIRnet): The joint SED and DOA estimation performance of SELDnet is compared with the method proposed by Hirvonen [20], hereafter referred to as HIRnet. The HIRnet was proposed for a circular array of omnidirectional microphones, hence we compare its performance only on the CANSYN and CRESYN datasets. HIRnet is a CNN-based network that uses the log-spectral power of each channel as the input feature and maps it to eight angles in full azimuth for each of the two classes (speech and music) as a multi-label classification task. More details about HIRnet can be found in [20]. In order to have a direct comparison with SELDnet-azi, we extend HIRnet to estimate DOAs at a $10°$ resolution for each of the sound event classes in our testing datasets.

C. Evaluation metrics

The proposed SELDnet is evaluated using individual metrics for SED and DOA estimation. For SED, we use the standard polyphonic SED metrics, error rate (ER) and F-score calculated in segments of one second with no overlap as proposed in [67, 68]. The segment-wise results are obtained from the frame-level predictions of the classifier by considering the sound events to be active in the entire segment if it is active in any of the frames within the segment. Similarly, we obtain labels for one-second segments of reference from its frame-wise annotation, and calculate the segment-wise ER and F-scores. Mathematically, the F-score is calculated as follows:

$$F = \frac{2 \cdot \sum_{k=1}^{K} TP(k)}{2 \cdot \sum_{k=1}^{K} TP(k) + \sum_{k=1}^{K} FP(k) + \sum_{k=1}^{K} FN(k)}, \tag{1}$$

where the number of true positives $TP(k)$ is the total number of sound event classes that were active in both reference and predictions for the kth one-second segment. The number of false positives in a segment $FP(k)$ is the number of sound event classes that were active in the prediction but were inactive in the reference. Similarly, $FN(k)$ is the number of false negatives, i.e. the number of sound event classes inactive in the predictions but active in the reference.

The ER metric is calculated as

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)}, \tag{2}$$

where, for each one-second segment k, $N(k)$ is the total number of active sound event classes in the reference. Substitution $S(k)$ is the number of times an event was detected but given the wrong level, this is obtained by merging the false negatives and false positives without individually correlating which false positive substitutes which false negative. The remaining false positives and false negatives, if any, are counted as insertions $I(k)$ and deletions $D(k)$ respectively. These statistics are mathematically defined as follows:

$$S(k) = \min(FN(k), FP(k)), \tag{3}$$

$$D(k) = \max(0, FN(k) - FP(k)), \tag{4}$$

$$I(k) = \max(0, FP(k) - FN(k)). \tag{5}$$

An SED method is jointly evaluated using the F-score and ER metric, and an ideal method will have an F-score of one (reported as percentages in Table) and ER of zero. More details regarding the F-score and ER metric can be read in [67, 68].

The predicted DOA estimates (x_E, y_E, z_E) are evaluated with respect to the reference (x_G, y_G, z_G) used to synthesize the dataset, utilizing the central angle $\sigma \in [0, 180]$. The σ is the angle formed by (x_E, y_E, z_E) and (x_G, y_G, z_G) at the origin in degrees, and is given by

$$\sigma = 2 \cdot \arcsin\left(\frac{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}{2}\right) \cdot \frac{180}{\pi}, \tag{6}$$

where, $\Delta x = x_G - x_E$, $\Delta y = y_G - y_E$, and $\Delta z = z_G - z_E$. The DOA error for the entire dataset is then calculated as

$$DOA\,error = \frac{1}{D} \cdot \sum_{d=1}^{D} \sigma((x_G^d, y_G^d, z_G^d), (x_E^d, y_E^d, z_E^d)) \tag{7}$$

where D is the total number of DOA estimates across the entire dataset, and $\sigma((x_G^d, y_G^d, z_G^d), (x_E^d, y_E^d, z_E^d))$ is the angle between d-th estimated and reference DOAs.

Additionally, in order to account for time frames where the number of estimated and reference DOAs are unequal, we report the frame recall, calculated as $TP/(TP + FN)$ in percentage, where true positives TP is the total number of time frames in which the number of DOAs predicted is equal to reference, and false negatives FN is the total number of frames where the predicted and reference DOA are unequal.

The DOA estimation method is jointly evaluated using the DOA error and the frame recall, and an ideal method has a frame recall of one (reported as percentages in Table) and DOA error of zero.

During the training of SELDnet, we perform early stopping based on the combined SELD score calculated as

$$SELD\,score = (SED\,score + DOA\,score)/2, \tag{8}$$

where

$$SED\,score = (ER + (1 - F))/2, \tag{9}$$

$$DOA\,score = \big(DOA\,error/180 + (1 - frame\,recall)\big)/2, \tag{10}$$

and an ideal SELD method will have an SELD score of zero. In the proposed method, the localization performance is dependent on the detection performance. This relation is represented by the frame recall metric of DOA estimation. As a consequence, the SELD score which is comprised of frame recall metric in addition to the SED metrics can be seen to weigh the SED performance more than DOA.

D. Experiments

The SELDnet is evaluated in different dimensions to understand its potential and drawbacks. The experiments carried out with different datasets in this regard are explained below.

1) SELDnet architecture and model parameter tuning: A wide variety of architectures with different combinations of CNN, RNN and FC layers are explored on the ANSYN $O2$ subset with frame length $M = 1024$ ($23.2\,ms$). Additionally, for each architecture, we tune the model parameters such as the number of CNN, RNN, and FC layers (0 to 4) and nodes (in the set of $[16, 32, 64, 128, 256, 512]$). The input sequence length is tuned in the set of $[32, 64, 128, 256, 512]$, the DOA and SED branch output loss weights in the set of $[1, 5, 50, 500]$, the regularization (dropout in the set of $[0, 0.1, 0.2, 0.3, 0.4, 0.5]$, L1 and L2 in the set of $[0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}]$) and the CNN max-pooling in the set of $[2, 4, 6, 8, 16]$ for each layer. The best set of parameters are the ones which give the lowest SELD score on the three cross-validation splits of the dataset. After finding the best network architecture and configuration, we tune the input audio feature parameter M by varying it in the set of $[512, 1024, 2048]$. Simultaneously the sequence length is also changed with respect to M such that the input audio length is kept constant ($1.49\,s$ obtained from the first round of tuning). We perform fine-tuning of model parameters for different M and sequence length values, this time only the number of CNN, RNN and FC nodes are tuned in a small range (neighboring nodes in the set of [16, 32, 64, 128, 256, 512]) to identify the optimum parameters. Similar fine-tuning is repeated for other datasets.

2) Selecting SELDnet output format: The output format for polyphonic SED in the literature has become standardized to estimating the temporal activity of each sound class using frame-wise binary numbers [31–34]. On the other hand, the output formats for DOA estimation are still being experimented with as seen in Table I. Among the DOA estimation methods using regression mode, there are two possible output formats, predicting azimuth and elevation, and predicting x, y, z coordinates of the DOA on the unit sphere. In order to identify the best output format among these two, we evaluate the SELDnet for both. During this evaluation, only the output weight parameter of the model is fine-tuned in the set of $[1, 5, 50, 500]$. Additionally, for a regression-based model, the default output i.e. the DOA target when the event is not active should be chosen carefully. In this study, we chose the default DOA output to be $180°$ in azimuth and $60°$ in elevation (the datasets do not contain sound events for these DOA values), and $x = 0$, $y = 0$ and $z = 0$ for default Cartesian outputs. The chosen default Cartesian coordinates are equidistant from all the possible DOA values. On the other hand, there are no such equidistant azimuth and elevation values. Hence we chose the default values ($180°$, $60°$) to be in a similar range as the true DOA values.

3) Continuous DOA estimation and performance on unseen DOA values: Theoretically, the advantage of using a regression-based DOA estimator over a classification-based one is that the network is not limited to a set of DOA angles, but it can operate as a high-resolution continuous DOA estimator. To study this, we train the SELDnet on ANSYN subsets whose sound events are placed on an angular grid of $10°$ resolution along azimuth and elevation, and test the model on a dataset where the angular grid is shifted by $5°$ along azimuth and elevation while keeping the temporal location unchanged. This shift makes the DOA values of the testing split unseen, and correctly predicting the DOAs will prove that the regression model can estimate the DOAs in a continuous space. Additionally, it also proves the robustness of the SELDnet to predict unseen DOA values.

4) Performance on mismatched reverberant dataset: Parametric DOA estimation methods are known to be sensitive to reverberation [48]. In this regard, we first evaluate the performance of SELDnet on the simulated (RESYN), and real-life (REAL, REALBIG, and REALBIGAMB) reverberant datasets and further compare the results with the parametric baseline MUSIC.

DNN based methods are known to fail when the training and testing splits come from different domains. For example, the performance of a DNN trained on anechoic dataset would be poor on a reverberant testing dataset. This performance can only be improved by training the DNN on a similar reverberant dataset as the testing dataset. On the other hand, it is impractical to train such a DNN for every existing room-dimension, its surface material distribution, and the reverberation times associated with it. In this regard, it would be ideal if the proposed method is robust to a moderate mismatch in reverberant conditions so that a single model can be used for a range of comparable room configurations. Motivated by this, we study the sensitivity of SELDnet on moderately mismatched reverberant data. Specifically, we train the SELDnet with RESYN room 1 dataset and test it on RESYN room 2 and 3 datasets that are mismatched in terms of volume and reverberation times as described in Section III-A2.

5) Performance on the size of the dataset: We study the performance of SELDnet on two datasets, REAL, and REALBIG that are similar in content, but different in size.

6) Performance with ambiance at different SNR: The performance of SELDnet with respect to different SNRs (0, 10 and 20 dB) of the sound event is studied on the REALBIGAMB dataset.

7) Generic to array structure: SELDnet is a generic method that learns to localize and recognize sound events from any array structure. This additionally implies that the SELDnet will continue to work in the desired manner if the configuration of the array such as individual microphone response, microphone spacing and the number of microphones remains the same between the training and testing set. If the array configuration changes between the training and testing set, then the SELDnet will have to be retrained for the new array configuration.

In order to prove that the SELDnet is applicable to any array configuration and not just dependent on the Ambisonics format, SELDnet is evaluated on a circular array. In comparison to the Ambisonic format, the chosen circular array has a different number of microphones, each placed on a single plane, and with an omnidirectional response. Further, we compare the SELDnet performance with dataset compatible

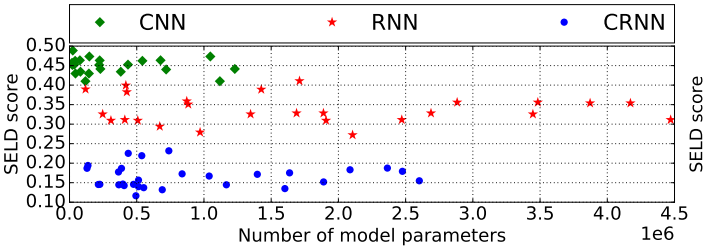


Fig. 3. SELD score for ANSYN $O2$ dataset for different CNN, RNN and CRNN architecture configurations.

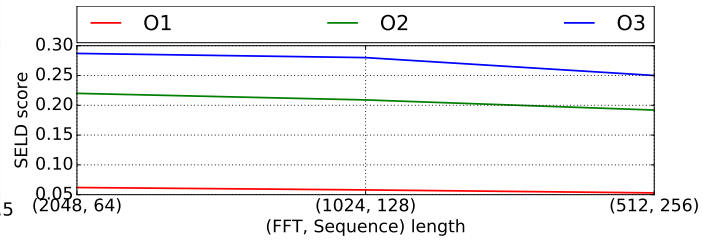


Fig. 4. SELD score for ANSYN datasets for different combinations of FFT length and input sequence length in frames.

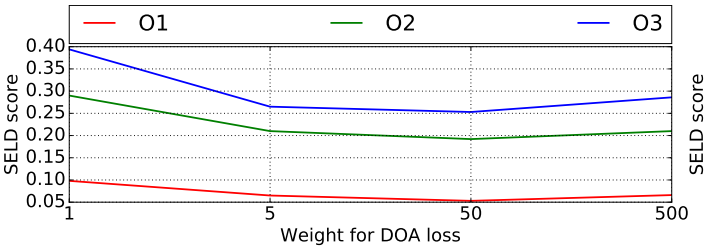


Fig. 5. SELD score for ANSYN datasets with respect to different weights for DOA output.

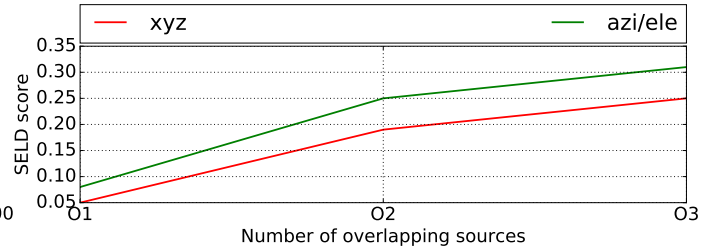


Fig. 6. SELD score for ANSYN datasets with respect to DOA output formats.

baselines such as SEDnet, MSEDnet, HIRnet, and AZInet. Since the HIRnet and AZInet baselines methods are proposed for estimating azimuth only, we compare the results with the SELDnet-azi version. Additionally, we also report the performance of using SELDnet with DOA estimation in x, y, z axis on CANSYN and CRESYN datasets.

In general, for all our experiments the only difference between the training and testing splits is the mutually exclusive set of sound examples. Apart from experiment III-D3 the training and testing splits contains the same set of spatial locations i.e., azimuth and elevation angles at 10° resolution amounting to 468 spatial locations (= 36 azimuth angles * 13 elevation angles). But the distance of the sound event at each of this 468 spatial locations is an added variable. For example, in the anechoic case, a sound event can be placed anywhere between 1-10 m at 0.5 m resolution. This variable amounts to 8892 spatial locations (= 468 * 19 distance positions) that are being coarsely grouped to 468 locations. This complexity is stretched further in experiment III-D3 where the testing split sound event examples and their spatial locations both are different from the training split.

IV. Results and Discussion

1) SELDnet architecture and model parameter tuning: The SELD scores obtained with hyper-parameter tuning of different CNN, RNN, and CRNN configurations as explained in Section III-D1 are visualized with respect to the number of model parameters in Figure 3. CNN in this figure refers to a SELDnet architecture which had no RNN layers but just CNN and FC layers. Similarly, RNN refers to SELDnet without CNN layers, while CRNN refers to SELDnet with CNN, RNN and FC layers. This experiment was carried out on ANSYN $O2$ dataset. The CRNN architecture was seen to perform the best followed by the RNN architecture.

The optimum model parameters across the ANSYN subsets after hyper-parameter tuning the CRNN architecture was found to have three layers of CNN with 64 nodes each (P in Figure 1a), followed by two layers of GRU with 128 nodes each (Q in Figure 1a), and one FC layer with 128 nodes (R in Figure 1a). The max-pooling over frequency after each of the three CNN layers (MP_i in Figure 1a) was $(8, 8, 2)$. This configuration had about 513,000 parameters.

Further, the SELDnet was seen to perform best with no regularization (dropout, or L1 or L2 regularization of weights). A frame length of $M = 512$ and sequence length of 256 frames was seen to give the best results across ANSYN subsets (Figure 4). Furthermore, on tuning the sequence length with frame length fixed ($M = 512$), the best scores were obtained using 512 frames ($2.97\,s$). Sequences longer than this could not be studied due to hardware restrictions. For the output weights, DOA output weighted 50 times more than SED output was seen to give the best results across subsets (Figure 5).

On fine-tuning the SELDnet parameters obtained with ANSYN dataset for RESYN subsets, the only parameter that helped improve the performance was using a sequence length of 256 instead of 512, leaving the total number of network parameters unchanged at 513,000. Similar configuration gave the best results for CANSYN and CRESYN datasets.

Model parameters identical to ANSYN dataset were observed to perform the best on the REAL subsets. The same parameters were also used for the study of REALBIG and REALBIGAMB subsets.

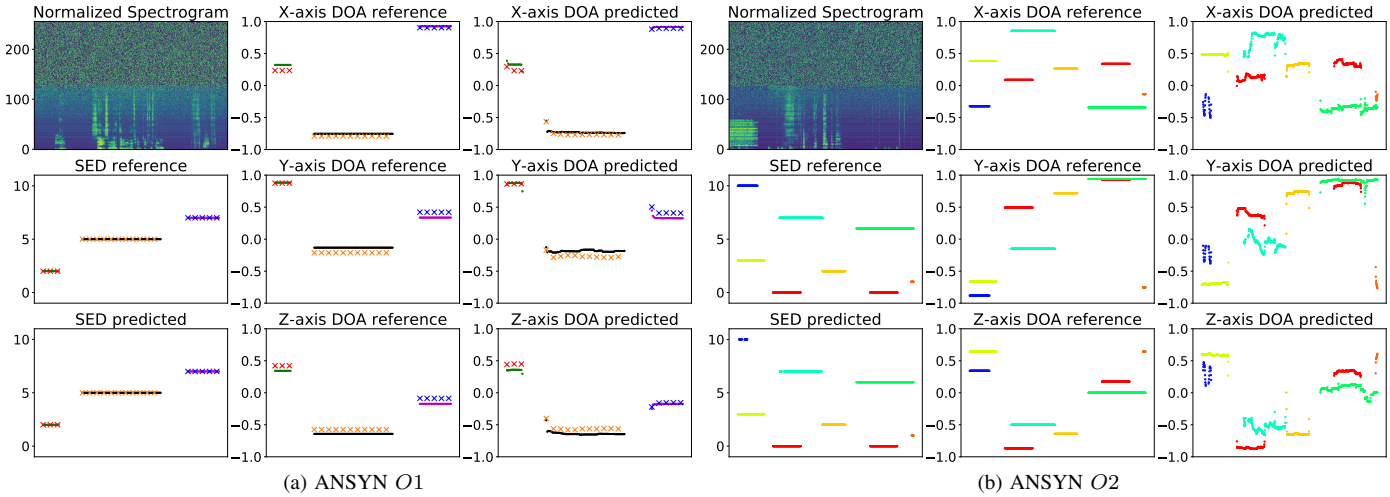
2) Selecting SELDnet output format: In the output data formats study, it was observed that using the Cartesian x, y, z format in place of azimuth/elevation angle was truly helping the network learn better across datasets as seen in Figure 6. This suggests that the discontinuity at the angle wrap-around boundary actually reduces the performance of DOA estimation and hence the SELD score.

TABLE IV
SED AND DOA ESTIMATION METRICS FOR ANSYN AND RESYN DATASETS. THE RESULTS FOR THE RESYN ROOM 2 AND 3 TESTING SPLITS WERE OBTAINED FROM CLASSIFIERS TRAINED ON RESYN ROOM 1 TRAINING SET. BEST SCORES FOR SUBSETS IN BOLD.

		ANSYN			RESYN Room 1			RESYN Room 2			RESYN Room 3		
	Overlap	1	2	3	1	2	3	1	2	3	1	2	3
SED metrics													
SELDnet	ER	**0.04**	0.16	0.19	**0.10**	0.29	0.32	**0.11**	0.33	0.35	**0.13**	0.32	0.34
	F	**97.7**	89.0	85.6	**92.5**	**79.6**	**76.5**	**91.6**	**79.5**	**75.8**	**89.8**	79.1	75.5
MSEDnet [39]	ER	0.10	**0.13**	**0.17**	0.17	**0.28**	**0.29**	0.19	**0.30**	**0.26**	0.18	**0.29**	**0.30**
	F	94.4	**90.1**	**87.2**	89.1	79.1	75.6	88.3	78.2	74.2	86.5	**80.5**	**76.1**
SEDnet [39]	ER	0.14	0.16	0.18	0.18	**0.28**	0.30	0.19	0.32	0.28	0.21	0.32	0.33
	F	91.9	89.1	86.7	88.2	76.9	74.1	87.6	76.4	73.2	85.1	78.2	75.6
DOA metrics													
SELDnet	DOA error	3.4	13.8	17.3	9.2	20.2	**26.0**	11.5	26.0	**33.1**	12.1	25.4	**31.9**
	Frame recall	**99.4**	**85.6**	**70.2**	**95.8**	**74.9**	**56.4**	**96.2**	**78.9**	**61.2**	**95.9**	**78.2**	**60.7**
DOAnet [25]	DOA error	**0.6**	8.0	18.3	**6.3**	**11.5**	38.4	**3.4**	**6.9**	-	**4.6**	**10.9**	-
	Frame recall	95.4	42.7	1.8	59.3	15.8	1.2	46.2	14.3	-	49.7	14.1	-
MUSIC	DOA error	4.1	**7.2**	**15.8**	40.2	47.1	50.5	45.7	58.1	74.0	48.3	60.6	75.6

3) Continuous DOA estimation and performance on unseen DOA values: The input and outputs of SELDnet trained on ANSYN $O1$ and $O2$ subsets for a respective 1000 frame test sequence are visualized in Figure 7. The Figure represents each sound event class and its associated DOA outputs with a unique color. In the case of ANSYN $O1$, we see that the network predictions of SED and DOA are almost perfect. In the case of unseen DOA values (× markers), the network predictions continue to be accurate. This shows that the regression mode output format helps the network learn continuous DOA values, and further that the network is robust to unseen DOA values. In case of ANSYN $O2$, the SED predictions are accurate, while the DOA estimates, in general, are seen to vary around the respective mean reference value. In this work, the DOA and SED labels for a single sound event instance are considered to be constant for the entire duration even though the instance has inherent magnitude variations and silences within. From Figure 7b it seems that these variations and silences are leading to fluctuating DOA estimates, while the SED predictions are unaffected. In general, we see that the proposed method successfully recognizes, localizes in time and space, and tracks multiple overlapping sound events simultaneously.

Table IV presents the evaluation metric scores for the SELDnet and the baseline methods with ANSYN and RESYN datasets. In the SED metrics for the ANSYN datasets, the SELDnet performed better than the best baseline MSEDnet for $O1$ subset while MSEDnet performed slightly better for $O2$ and $O3$ subsets. With regard to DOA metrics, the SELDnet is significantly better than the baseline DOAnet in terms of frame recall. This improvement in frame recall is a direct result of using SED output as a confidence measure for estimating DOA, thereby extending state-of-the-art SED performance to SELD. Although the frame recall of DOAnet is poor, its DOA error for $O1$ and $O2$ subsets is observed to be lower than SELDnet. The DOA error of the parametric baseline MUSIC with the knowledge of the number of sources is seen to be the best among the evaluated methods for $O2$ and $O3$ subsets.



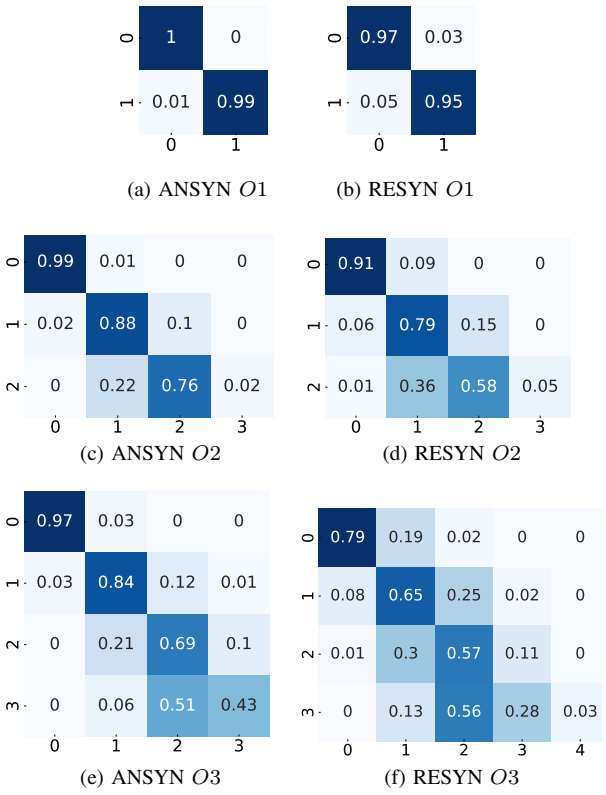
(a) ANSYN $O1$ (b) ANSYN $O2$

Fig. 7. SELDnet input and outputs visualized for ANSYN $O1$ and $O2$ datasets. The horizontal-axis of all sub-plots for a given dataset represents the same time frames, the vertical-axis for spectrogram sub-plot represents the frequency bins, vertical-axis for SED reference and prediction sub-plots represents the unique sound event class identifier, and for the DOA reference and prediction sub-plots, it represents the distance from the origin along the respective axes. The bold lines visualize both the reference labels and predictions of DOA and SED for ANSYN $O1$ and $O2$ datasets, while the × markers in Figure 7a visualize the results for testing split with unseen DOA values (shifted by 5° along azimuth and elevation).

TABLE V
SED AND DOA ESTIMATION METRICS FOR REAL, REALBIG AND REALBIGAMB DATASETS. BEST SCORES FOR SUBSETS IN BOLD.

		REAL			REALBIG			REALBIGAMB 20dB			REALBIGAMB 10dB			REALBIGAMB 0dB		
	Overlap	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
SED metrics																
SELDnet	ER	0.40	0.49	0.53	0.37	0.42	0.50	**0.34**	0.46	0.52	**0.37**	0.49	0.52	**0.46**	0.58	0.59
	F	60.3	53.1	51.1	65.4	61.5	56.5	65.6	58.5	55.0	**66.3**	55.4	53.3	**57.9**	48.6	49.0
MSEDnet [39]	ER	**0.35**	**0.38**	**0.41**	**0.34**	**0.39**	**0.38**	0.35	**0.40**	**0.41**	0.38	**0.43**	**0.42**	0.48	**0.56**	**0.54**
	F	**66.2**	**61.6**	**59.5**	67.3	61.8	61.9	**66.0**	**61.6**	**60.1**	63.2	**58.7**	**59.3**	54.5	**49.3**	**51.3**
SEDnet [39]	ER	0.38	0.42	0.43	0.38	0.43	0.44	0.39	0.42	0.43	0.41	0.44	0.46	0.51	0.61	0.57
	F	64.6	61.5	57.2	**68.0**	**62.4**	**62.4**	65.7	60.1	59.2	62.7	56.3	56.9	52.6	46.0	50.4
DOA metrics																
SELDnet	DOA error	26.6	33.7	36.1	23.1	31.3	34.9	25.4	32.5	36.1	27.2	32.5	36.1	30.7	33.7	36.7
	Frame recall	**64.9**	**41.5**	**24.6**	**68.0**	**45.2**	**28.3**	**69.1**	**42.8**	**25.8**	**66.9**	**40.0**	**27.3**	**62.5**	**35.2**	**23.4**
DOAnet [25]	DOA error	**6.3**	**20.1**	**25.8**	**7.5**	**17.8**	**22.9**	**6.3**	**18.9**	**25.78**	**8.0**	**20.1**	**24.1**	**14.3**	**24.1**	**27.5**
	Frame recall	46.5	11.5	2.9	44.1	12.5	3.1	34.7	11.6	3.2	42.1	13.5	3.3	30.1	10.5	2.8
MUSIC	DOA error	36.3	49.5	54.3	35.8	49.6	53.8	54.5	56.1	61.3	51.6	54.5	62.6	41.9	47.5	62.3

4) Performance on mismatched reverberant dataset: From Table IV results on RESYN room 1 subsets, we see that the performance of parametric method MUSIC is poor in comparison to SELDnet in reverberant conditions. The SELDnet is seen to perform significantly better than the baseline DOAnet in terms of frame recall, although the DOAnet has lower DOA error for $O1$ and $O2$ subsets. The SED metrics of SELDnet are comparable if not better than the best baseline performance of MSEDnet. Further, on training the SELDnet on room 1 dataset and testing on moderately mismatched reverberant room 2 and 3 datasets the SED and DOA metric trends remain similar to the results of room 1 testing split. That is, the SELDnet has higher frame recall, the DOAnet has better DOA error, the MUSIC performs poorly, and the SED metrics of SELDnet are comparable to MSEDnet. These results prove that the SELDnet is robust to reverberation in comparison to baseline methods and performs seamlessly on moderately mismatched room configurations.



(a) ANSYN $O1$ (b) RESYN $O1$

(c) ANSYN $O2$ (d) RESYN $O2$

(e) ANSYN $O3$ (f) RESYN $O3$

Fig. 8. Confusion matrix for the number of sound event classes estimated to be active per frame by the SELDnet for ANSYN and RESYN datasets. The horizontal axis represents the SELDnet estimate, and the vertical axis represents the reference.

Figure 8 visualizes the confusion matrices for the estimated number of sound event classes per frame by SELDnet. For example in Figure 8c the SELDnet correctly estimated the number of sources to be two in 76% (true positive percentage) of the frames which had two sources in the reference. In context, the frame recall value used as a metric to evaluate DOA estimation represents this confusion matrix in one number. From the confusion matrices, we observe that the percentage of true positives drops with higher number of sources, and this drop is even more significant in the reverberant scenario. But, in comparison to the frame recall metric of the baseline DOAnet in Table IV, the SELDnet frame recall is significantly better for higher number of overlapping sound events, especially in the reverberant conditions.

5) Performance on the size of the dataset: The overall performance of SELDnet on REAL dataset (Table V) reduced in comparison to ANSYN and RESYN datasets. The baseline MSEDnet is seen to perform better than SELDnet in terms of SED metrics. Similar performance drop on real-life datasets has also been reported on SED datasets in other studies [37]. With regard to DOA metrics, the frame recall of SELDnet continues to be significantly better than DOAnet, while the DOA error of DOAnet is lower than SELDnet. The performance of MUSIC is seen to be poor in comparison to both DOAnet and SELDnet. With the larger REALBIG dataset the SELDnet performance was seen to improve. A similar study was done with larger ANSYN and RESYN datasets, where the results were comparable with that of smaller datasets. This shows that the datasets with real-life IR are more complicated than synthetic IR datasets, and having larger real-life datasets helps the network learn better.

6) Performance with ambiance at different SNR: In presence of ambiance, SELDnet was seen to be robust for 10 and 20 dB SNR REALBIGAMB datasets (Table V). In comparison to the SED metrics of REALBIG dataset with no ambiance, the SELDnet performance on $O1$ subsets of 10 dB and 20 dB

TABLE VI
SED AND DOA ESTIMATION METRICS FOR CANSYN AND CRESYN DATASETS. BEST SCORES FOR SUBSETS IN BOLD.

		CANSYN			CRESYN		
	Overlap	1	2	3	1	2	3
SED metrics							
SELDnet	ER	0.11	**0.18**	0.19	0.13	0.22	0.30
	F score	93.0	86.6	85.3	90.4	82.2	78.0
SELDnet-azi	ER	**0.08**	0.19	0.24	**0.06**	**0.18**	**0.20**
	F score	**94.7**	87.5	83.8	**96.3**	**87.9**	**85.6**
MSEDnet [39]	ER	0.09	**0.18**	**0.16**	0.12	0.22	0.26
	F score	94.6	**89.0**	**86.7**	92.7	83.7	80.7
SEDnet [39]	ER	0.15	0.21	0.20	0.18	0.26	0.25
	F score	91.4	87.3	84.7	90.5	84.3	82.8
HIRnet [20]	ER	0.41	0.45	0.62	0.43	0.46	0.50
	F score	60.0	54.9	58.8	59.3	60.2	58.6
DOA metrics							
SELDnet	DOA error	29.5	31.3	34.3	28.4	33.7	41.0
	Frame recall	97.9	78.8	67.0	96.4	75.7	60.7
SELDnet-azi	DOA error	7.5	14.4	19.6	5.2	13.2	18.4
	Frame recall	98.0	**82.1**	**66.2**	**98.5**	**82.3**	**70.6**
HIRnet [20]	DOA error	5.2	16.3	33.0	7.4	18.6	43.3
	Frame recall	60.2	35.9	18.4	56.9	20.5	10.7
AZInet [18]	DOA error	**1.2**	**4.0**	7.4	**2.3**	**6.9**	**9.7**
	Frame recall	**99.4**	80.5	60.5	97.3	65.2	44.8

ambiance is comparable, while a small drop in performance was observed with the respective $O2$ and $O3$ subsets. Whereas, the performance was seen to drop considerably for the 0 dB SNR dataset. With respect to DOA error, the SELDnet performed better than MUSIC but poorer than DOAnet across datasets, on the other hand, SELDnet gave significantly higher frame recall than DOAnet. From the insight of SELDnet performance on REAL dataset (Section IV-5), the more complex the acoustic scene the larger the dataset size required to learn better. Considering that the SELDnet is jointly estimating the DOA along with SED in a challenging acoustic scene with ambiance the SELDnet performance can potentially improve with larger datasets.

7) Generic to array structure: The results on circular array datasets are presented in Table VI. With respect to SED metrics, the SELDnet-azi performance is seen to be better than the best baseline MSEDnet for all subsets of CRESYN dataset, while MSEDnet is seen to perform better for $O2$ and $O3$ subsets of CANSYN dataset. Similarly, in the case of DOA metrics, the SELDnet-azi has better frame recall than the best baseline method AZInet across datasets (except for CANSYN $O1$). Whereas, AZInet has lower DOA error than SELDnet-azi. Between SELDnet and SELDnet-azi, even though the frame recall is in the same order the DOA error of SELDnet-azi are lower than SELDnet. This shows that estimating DOA in 3D (x, y, z) is challenging using a circular array. Overall, the SELDnet is shown to perform consistently across different array structures (Ambisonic and circular array), with good results in comparison to baselines.

The usage of SED output as a confidence measure for estimating the number of DOAs in the proposed SELDnet is shown to improve the frame recall significantly and consistently across the evaluated datasets. On the other hand, the DOA error obtained with SELDnet is consistently higher than the classification based baseline DOA estimation methods [18, 25]. We believe that this might be a result of the regression-based DOA estimation approach in SELDnet not having completely learned the full mapping between input feature and the continuous DOA space. The investigation of which is planned for future work. In general, a classification only or a classification-regression based SELD approach can be chosen based on the required frame recall, DOA error, resolution of DOA labels, training split size, and robustness to unseen DOA values and reverberation.

V. CONCLUSION

In this paper, we proposed a convolutional recurrent neural network (SELDnet) to simultaneously recognize, localize and track sound events with respect to time. The localization is done by estimating the direction of arrival (DOA) on a unit sphere around the microphone using 3D Cartesian coordinates. We tie each sound event output class in the SELDnet to three regressors to estimate the respective Cartesian coordinates. We show that using regression helps estimating DOA in a continuous space, and also estimating unseen DOA values accurately. On the other hand, estimating a single DOA for each sound event class does not allow recognizing multiple instances of the same class overlapping. We plan to tackle this problem in our future work.

The usage of SED output as a confidence measure to estimate DOA was seen to extend the state-of-the-art SED performance to SELD resulting in a higher recall of DOAs. With respect to the estimated DOA error, although the classification based baseline methods had poor recall they had lower DOA error in comparison to the proposed regression based DOA estimation. The proposed SELDnet uses phase and magnitude spectrogram as the input feature. The usage of such non-method-specific feature makes the method generic and easily extendable to different array structures. We prove this by evaluating on datasets of Ambisonic and circular array format. The proposed SELDnet is shown to be robust to reverberation, low SNR scenarios and unseen rooms with comparable room-sizes. Finally, the overall performance on dataset synthesized using real-life impulse response (IR) was seen to drop in comparison to artificial IR dataset, suggesting the need for larger real-life training datasets and more powerful classifiers in future.

REFERENCES

[1] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[2] ——, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016.

[3] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," in *Journal of Robotics and Mechatronics*, vol. 29, no. 1, 2017.

[4] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *International Conference on Robotics and Automation (ICRA)*, 2018.

[5] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," in *ACM Computing Surveys (CSUR)*, 2016.

[6] C. Grobler, C. Kruger, B. Silva, and G. Hancke, "Sound based localization and identification in industrial environments," in *IEEE Industrial Electronics Society (IECON)*, 2017.

[7] P. W. Wessels, J. V. Sande, and F. V. der Eerden, "Detection and localization of impulsive sound events for environmental noise assessment," in *The Journal of the Acoustical Society of America 141*, vol. 141, no. 5, 2017.

[8] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, 2015.

[9] C. Busso, S. Hernanz, C.-W. Chu, S.-i. Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan, "Smart room: participant and speaker localization and identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[10] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

[11] M. Wölfel and J. McDonough, *Distant speech recognition*. John Wiley & Sons, 2009.

[12] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," in *IEEE Signal Processing Letters*, vol. 21, 2014.

[13] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *European Signal Processing Conference (EUSIPCO)*, 2011.

[14] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time-frequency audio features," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, 2009.

[15] T. A. Marques *et al.*, "Estimating animal population density using passive acoustics," in *Biological reviews of the Cambridge Philosophical Society*, vol. 88, no. 2, 2012.

[16] B. J. Furnas and R. L. Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," in *Journal of Wildlife Management*, vol. 79, no. 2, 2014.

[17] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[18] ——, "Multi-speaker localization using convolutional neural network trained with noise," in *Neural Information Processing Systems (NIPS)*, 2017.

[19] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[20] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*, 2015.

[21] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," in *International Journal of Applied Engineering Research*, vol. 12, no. 22, 2017.

[22] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[23] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[24] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilisitic neural network," in *IEEE Transactions on Industrial Electronics*, vol. 29, no. 1, 2017.

[25] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference (EUSIPCO)*, 2018.

[26] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Deutsche Jahrestagung für Akustik (DAGA)*, 2015.

[27] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *European Signal Processing Conference (EUSIPCO)*, 2010.

[28] E. Çakır, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi-label deep neural networks," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2015.

[29] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[30] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[31] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, and K. Takeda, "Duration-controlled LSTM for polyphonic sound event detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, 2017.

[32] M. Zöhrer and F. Pernkopf, "Virtual adversarial training and data augmentation for acoustic event detection with gated recurrent neural networks," in *INTERSPEECH*, 2017.

[33] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[34] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," in *INTERSPEECH*, 2016.

[35] S. Adavanne, A. Politis, and T. Virtanen, "Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2018.

[36] H. Lim, J. Park, K. Lee, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[37] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, 2017.

[38] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[39] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[40] I.-Y. Jeong, S. Lee1, Y. Han, and K. Lee, "Audio event detection using multiple-input convolutional neural network," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[41] J. Zhou, "Sound event detection in multichannel audio LSTM network," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[42] R. Lu and Z. Duan, "Bidirectional gru for sound event detection," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[43] A. Temko, C. Nadeu, and J.-I. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07," in *Multimodal Technologies for Perception of Humans*. Springer, 2008.

[44] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," in *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, 2001.

[45] M. S. Brandstein and H. F. Silverman, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

[46] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, 1986.

[47] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 37, no. 7, 1989.

[48] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Springer, 2001.

[49] H. Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*. Springer, 2007, vol. 348.

[50] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[51] J. Traa and P. Smaragdis, "Multiple speaker tracking with the Factorial Von Mises-Fisher filter," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014.

[52] R. Chakraborty and C. Nadeu, "Sound-model-based acoustic source localization using distributed microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[53] K. Lopatka, J. Kotus, and A. Czyzewsk, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools and Applications Journal*, vol. 75, no. 17, 2016.

[54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, 2015.

[55] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[56] F. Chollet, "Keras v2.0.8," 2015, accessed on 7 May 2018. [Online]. Available: https://github.com/fchollet/keras

[57] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, accessed on 7 May 2018. [Online]. Available: https://www.tensorflow.org/

[58] E. Benetos, M. Lagrange, and G. Lafay, "Sound event detection in synthetic audio," 2016, accessed on 7 May 2018. [Online]. Available: https://archive.org/details/dcase2016_task2_train_dev

[59] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," in *The Journal of the Acoustical Society of America*, vol. 65, no. 4, 1979.

[60] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009.

[61] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[62] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM International Conference on Multimedia (ACM-MM)*, 2014.

[63] A. Politis, "Microphone array processing for parametric spatial audio techniques," *Ph.D. thesis, Aalto University*, 2016.

[64] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: tasks, datasets and baseline system," in *Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE)*, 2017.

[65] B. Ottersten, M. Viberg, P. Stoica, and A. Nehorai, "Exact and large sample maximum likelihood techniques for parameter estimation and detection in array processing," in *Radar Array Processing. Springer Series in Information Sciences*, 1993.

[66] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, 2012.

[67] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, vol. 6, no. 6, 2016.

[68] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. Plumbley, and D. Ellis, Eds. Springer International Publishing, 2018, ch. 6.

Sharath Adavanne received his M.Sc. degree in Information Technology from Tampere University of Technology (TUT), Finland in 2011. From 2011 to 2016 he worked in the industry solving problems related to music information retrieval, speech recognition, audio fingerprinting and general audio content analysis. Since 2016, he is pursuing his Ph.D. degree at the laboratory of signal processing in TUT. His current research interest is in the application of machine learning based methods for real-life auditory scene analysis.

Archontis Politis obtained a M.Eng degree in civil engineering Aristotle University, Thessaloniki, Greece, and his M.Sc degree in sound and vibration studies from Institute of Sound and Vibration Studies (ISVR), Southampton, UK, in 2006 and 2008 respectively. From 2008 to 2010 he worked as a graduate acoustic consultant in Arup Acoustics, UK, and as a researcher in a joint collaboration between Arup Acoustics and the Glasgow School of Arts, on architectural auralization using spatial sound techniques. In 2016 he obtained a Doctor of Science degree on the topic of parametric spatial sound recording and reproduction from Aalto University, Finland. He has also completed an internship at the Audio and Acoustics Research Group of Microsoft Research, during summer of 2015. He is currently a post-doctoral researcher at Aalto University. His research interests include spatial audio technologies, virtual acoustics, array signal processing and acoustic scene analysis.

Joonas Nikunen received the M.Sc degree in signal processing and communications engineering and Ph.D degree in Signal Processing from Tampere University of Technology (TUT), Finland in 2010 and 2015, respectively. He is currently a post-doctoral researcher at TUT focusing on sound source separation with applications on spatial audio analysis, modification and synthesis. His other research interests include microphone array signal processing, 3D/360 audio in general and machine and deep learning for source separation.

Tuomas Virtanen is Professor at Laboratory of Signal Processing, Tampere University of Technology (TUT), Finland, where he is leading the Audio Research Group. He received the M.Sc. and Doctor of Science degrees in information technology from TUT in 2001 and 2006, respectively. He has also been working as a research associate at Cambridge University Engineering Department, UK. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition and music content analysis. Recently he has done significant contributions to sound event detection in everyday environments. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored more than 150 scientific publications on the above topics, which have been cited more than 6000 times. He has received the IEEE Signal Processing Society 2012 best paper award for his article "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria" as well as three other best paper awards. He is an IEEE Senior Member, member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society, Associate Editor of IEEE/ACM Transaction on Audio, Speech, and Language Processing, and recipient of the ERC 2014 Starting Grant.

Publication VI

Sharath Adavanne, Archontis Politis, Tuomas Virtanen, "Localization, Detection, and Tracking of Multiple Moving Sources with Convolutional Recurrent Neural Network," *Detection and Classification of Acoustic Scenes and Events (DCASE).* New York, USA, pp. 20-24, October 2019.

LOCALIZATION, DETECTION AND TRACKING OF MULTIPLE MOVING SOUND SOURCES WITH A CONVOLUTIONAL RECURRENT NEURAL NETWORK

Sharath Adavanne, Archontis Politis, and Tuomas Virtanen

Audio Research Group, Tampere University, Finland, firstname.lastname@tuni.fi

ABSTRACT

This paper investigates the joint localization, detection, and tracking of sound events using a convolutional recurrent neural network (CRNN). We use a CRNN previously proposed for the localization and detection of stationary sources, and show that the recurrent layers enable the spatial tracking of moving sources when trained with dynamic scenes. The tracking performance of the CRNN is compared with a stand-alone tracking method that combines a multi-source direction of arrival estimator and a particle filter. Their respective performance is evaluated in various acoustic conditions such as anechoic and reverberant scenarios, stationary and moving sources at several angular velocities, and with a varying number of overlapping sources. The results show that the CRNN manages to track multiple sources more consistently than the parametric method across acoustic scenarios, but at the cost of higher localization error.

***Index Terms*—** Multiple object tracking, recurrent neural network, sound event detection, acoustic localization

1. INTRODUCTION

Sound event localization, detection, and tracking (SELDT) is the combined task of identifying the temporal onset and offset of potentially temporally-overlapping sound events, recognizing their classes, and tracking their respective spatial trajectory when they are active. Performing SELDT successfully provides an automatic description of the acoustic scene that can be employed by machines to interact naturally with their surroundings. Applications such as teleconferencing systems and robots can use this information for tracking the sound event of interest [1–6]. Furthermore, smart cities and smart homes can use it for audio surveillance [7–9].

The joint localization and detection in static scenes with spatially stationary sources have been studied with different parametric [5, 8, 10, 11] and deep neural network (DNN) [12] based methods. However, these methods do not employ any temporal modeling required for the tracking of moving sources in dynamic scenes. Recently, we proposed a convolutional recurrent neural network (SELDnet) that was shown to perform significantly better localization and detection than the only other existing DNN-based method [12]. SELDnet's capabilities to localize events in full azimuth and elevation under matched and unmatched acoustic conditions, and without relying on features dependent on specific microphone arrays, were studied and presented in [13]. However, all the existing DNN-based methods including[12, 13] have only studied static scenes.

On the other hand, stand-alone tracking methods have been widely studied for both stationary and moving sources based on spatial information only [14–20], additional spectral information [21, 22], or in conjunction with visual information [23]. Such parametric methods often require manual tuning of multiple parameters corresponding to the scene composition and dynamics, and new sets of parameters have to be identified manually for different sound scenes. Furthermore, tracking usually focuses on distinguishing source trajectories, with no regard to source signal content. In the case of temporally overlapping trajectories, track identities are assigned to individual trajectories, but these identities are not source dependent and are generally re-used for trajectories from different sources across the audio recording. A balance between consistent association and localization determines the tracker's performance in most cases. Alternatively, a detect-before-track approach, as in the proposed SELDnet, circumvents the association problem by first detecting the active sound events, and then assigning a track to each detected event. As long as such a system is able to react to time-varying conditions, with temporally and spatially overlapping sound events from both stationary and moving sources, it is also able to detect and track the sound events of interest.

In this work, we study the multi-source tracking capabilities of a detection and localization system based on our recently proposed SELDnet [13]. To the best of the authors knowledge, this is the first DNN-based SELDT studies. We show that training the SELDnet with dynamic scene data results in tracking, in addition to localization and detection. This tracking ability is enabled by the recurrent layers of the SELDnet that can model the evolution of spatial parameters as a sequence prediction task given the sequential features and their spatial trajectory information. We show that the recurrent layers are crucial for tracking, and in comparison to stand-alone trackers they additionally perform detection. Unlike the parametric tracking methods discussed earlier, the recurrent layer is a generic tracking method that learns directly from the data without manual tracker-engineering. Finally, we show that the tracking performance of SELDnet is comparable with stand-alone parametric tracking methods through evaluation on five datasets, representing scenarios with stationary and moving sources at different angular velocities, anechoic and reverberant environments, and different numbers of overlapping sources. The method and all the studied datasets are publicly available[1].

2. METHOD

The block diagram of SELDnet [13] is illustrated in Figure 1. The input to SELDnet is a multichannel audio recording, from which a feature extraction block extracts the phase and magnitude components of the spectrogram from each channel. The SELDnet maps the input spectrogram of T-frames length to two outputs of the same length – sound event detection, and tracking; together they

This work has received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND.

[1] https://github.com/sharathadavanne/seld-net

https://doi.org/10.33682/xb0q-a335

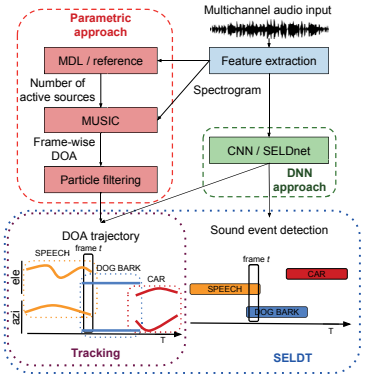


Figure 1: Workflows for the parametric tracking and DNN-based SELDT approaches. The sound class coloring and naming for the tracking task is only shown here to visualize the concept better. In practice tracking methods do not produce sound class labels as shown in Figure 3.

produce the SELDT output. The detection output is the class-wise probabilities for the C classes in the dataset of dimension $T \times C$, and is obtained as a multiclass multilabel classification task. The tracking output is a single direction of arrival (DOA) estimate per time frame for each sound class C as a multi-output regression task. Thus, when multiple instances of the same sound class are temporally overlapping, the SELDnet tracks only one instance or oscillate between the multiple instances. The estimated DOA is represented using 3D Cartesian coordinates of a point on a unit sphere around the microphone. The overall tracking output is of dimension $T \times 3C$, where $3C$ represents the three axes of the 3D Cartesian coordinates of a DOA for each class in C. Finally, to obtain the SELDT results, the class-wise probabilities of the detection output are binarized with a threshold of 0.5, anything greater represents the presence of the sound class and smaller represents the absence. The presence of a sound class in consecutive frames gives the onset and offset times, and the corresponding frame-wise DOA estimates from the tracking output when the sound class is active gives the DOA trajectory.

The SELDnet architecture used in this paper is identical to [13], with three convolutional layers of 64 filters each, followed by two-layers of 128-node gated recurrent units. The convolutional layers in the SELDnet are used as a feature extractor to produce robust features for detection and tracking. The recurrent layers are employed to model the temporal structure and the trajectory of the sound events. The output of the recurrent layers is shared between two branches of dense layers each with 128 units producing the detection and tracking estimates. The training and inference procedures of SELDnet are similar to [13] and is identical for both static and dynamic scenes, i.e., the same SELDnet designed for static scenes performs tracking when trained with moving scene data.

The recurrent layers utilize the current input frame along with the information learned from the previous input frames to produce the output for the current frame. This process is similar to a particle filter, which is a popular stand-alone parametric tracker and is also used as a baseline in this paper (see Section 3.3). The particle filter prediction at the current time frame is influenced by both the knowledge accumulated from the previous time frames and the input at the current time frame. For the tracking task of this paper, the particle filter requires the specific knowledge of the sound scene such as the spatial distribution of sound events, their respective velocity ranges

Table 1: Summary of Datasets

Sources	Sound scene	Impulse response	Acronym
Stationary [13]	Anechoic	Synthetic	ANSYN
	Reverberant		RESYN
		Real-life	REAL
Moving	Anechoic	Synthetic	MANSYN
	Reverberant	Real-life	MREAL

when active, and their probability of birth and death. Such concepts are not explicitly modeled in the recurrent layers used in SELDnet, rather they learn equivalent information directly from the input convolutional layer features and corresponding target outputs in the development dataset. In fact, recurrent layers have been shown to work as generic trackers [24] that can learn temporal associations of the target source from any sequential input features. Unlike the particle filters that only work with conceptual representations such as frame-wise multiple DOAs for tracking, the recurrent layers work seamlessly with both conceptual and latent representations such as convolutional layer features.

Finally, by training the recurrent layers in SELDnet using the loss calculated from both detection and tracking, the recurrent layers learn associations between DOAs from neighboring frames corresponding to the same sound class and hence produce the SELDT results. In general, unlike the parametric trackers, the recurrent layers perform similar tracking of the frame-wise DOAs in addition to also detecting their corresponding sound classes. Further, the recurrent layers do not need complicated problem-specific tracker- or feature-engineering that are required by the parametric trackers. A more theoretical relationship between recurrent layers and particle filter is presented in [25].

3. EVALUATION PROCEDURE

3.1. Datasets

The performance of SELDnet is evaluated on five datasets that are summarized in Table 1. We continue to use the stationary source datasets: ANSYN, RESYN and REAL from our previous work [13] to evaluate the tracking performance of the parametric tracker that was missing in [13], and compare with SELDnet. The recordings in ANSYN and RESYN are synthesized in anechoic and reverberant environments respectively. The recordings in REAL are synthesized by convolving isolated real-life sound events with real-life impulse responses collected at different spatial locations within a room. Further, we create moving-source versions of the ANSYN and REAL datasets, hereafter referred as MANYSYN and MREAL, to evaluate the performance on moving sources. The recordings of all datasets are 30 seconds long and captured in the four-channel first-order Ambisonics format [26]. Each dataset has three subsets with no temporally overlapping sources $O1$, maximum two $O2$, and maximum three temporally overlapping sources $O3$. Each of these subsets has three cross-validation splits consisting of 240 recordings for development and 60 for evaluation. All the synthetic impulse response datasets (ANSYN, RESYN and MANYSN) have sound events from 11 classes and DOAs with full azimuth range and elevation range $\in [-60, 60)$. The real-life impulse response datasets (REAL and MREAL) have 8 sound event classes and DOAs in full azimuth range and elevation range $\in [-40, 40)$. During the synthesis of stationary source datasets, all the sound events are placed in a spatial grid of 10° resolution for both azimuth and elevation angles. We refer the readers to [13] for more details on these datasets.

The anechoic moving source dataset MANSYN has the same sound event classes as ANSYN and is synthesized as follows. Every event is assigned a spatial trajectory on an arc with a constant distance from the microphone (in the range 1-10 m) and moving

with a constant angular velocity for its duration. Due to the choice of the ambisonic spatial recording format, the steering vectors for a plane wave source or point source in the far field are frequency-independent. Hence, there is no need for a time-variant convolution or impulse response interpolation scheme as the source is moving; the spatial encoding of the monophonic signal was done sample-by-sample using instantaneous ambisonic encoding vectors for the respective DOA of the moving source. The synthesized trajectories in MANSYN vary in both azimuth and elevation, and are simulated to have a constant angular velocity in the range $\in [-90^\circ, 90^\circ]/s$ with $10^\circ/s$ steps. Similarly, the MREAL dataset was synthesized with real-life impulse responses from [13] that were sampled at 1° resolution along azimuth only. Hence, unlike MANSYN, the sound events in MREAL (that are identical to REAL) have motion only along the azimuth with a constant angular velocity in the range $\in [-90^\circ, 90^\circ]/s$ and $10^\circ/s$ steps.

3.2. Metrics

The evaluation of the SELDT performance is done using individual metrics for detection and tracking identical to [13]. As the detection metric, we use the F-score and error rate calculated in segments of one-second with no overlap [27]. An ideal detection method will have an F-score of one and an error rate of zero. As the tracking metric, we use two frame-wise metrics: the frame recall and DOA error. The frame recall gives the percentage of frames in which the number of predicted DOAs is equal to the reference. The DOA error is calculated as the angle in degrees between the predicted and reference DOA. In order to associate multiple estimated DOAs with the reference, we use the Hungarian algorithm [28] to identify the smallest mean angular distance and use it as DOA error. An ideal tracking method has a frame recall of one and DOA error of zero (see [13] for more details).

3.3. Baseline Method

In the absence of publicly available implementations of multiple moving sound sources trackers, we use a combination of MUSIC [29] and an RBMCDA particle filter [30] to obtain tracking results in a similar fashion as in [15] and further made it publicly available [2]. The workflow of the baseline method is shown in Figure 1. MUSIC is a widely used [13, 31] subspace-based high-resolution DOA estimation method that can detect multiple narrowband sources. It relies on an eigendecomposition of the narrowband spatial covariance matrix computed from the multichannel spectrogram, and it additionally requires a source number estimate in order to distinguish between a signal and noise subspace. Herein, the number of active sources is taken from the reference of the dataset. To obtain broadband DOA estimates, the narrowband covariance matrices are averaged across three consecutive frames and frequency bins from 50 Hz to 8 kHz. We perform 2D spherical peak-finding on the resulting MUSIC pseudospectrum generated on a 2D angular grid with a 10° resolution for stationary and 1° for moving sources, in both azimuth and elevation. The final output of MUSIC MUS_{GT} is a list of frame-wise DOAs corresponding to the highest peaks equal to the number of active sources in each frame.

The second stage of the parametric method involves a particle filter that produces tracking results by processing the frame-wise DOA information of MUSIC MUS_{GT}. The particle filter assumes that the number of sources at each time frame is unknown and tracks them with respect to time using a fixed number of particles. At each time frame, the particle filter receives multiple DOAs and, based on knowledge accumulated from the previous time frames, it assigns each new DOA to one of the existing trajectories, clutter (noise), or a newborn source. Additionally, it also decides if any of the existing trajectories have died. The final output of the particle filter MUS^{PF}_{GT} produces the temporal onset-offset and the DOA trajectory for each of the active sound events. We refer the reader to [30] for the details of this approach.

[2] https://github.com/sharathadavanne/multiple-target-tracking

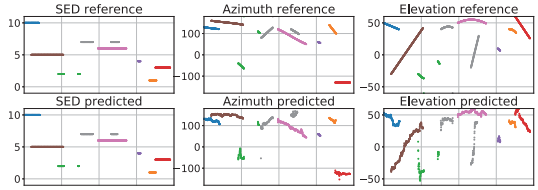


Figure 2: Visualization of the SELDnet predictions and its respective reference for a MANSYN $O2$ dataset recording. The horizontal-axis of all sub-plots represents the same time frames. The vertical-axis represents sound event class indices for the detection subplots, and DOA azimuth and elevation angles in degrees for remaining subplots.

3.4. Experiments

In all our experiments, the baseline particle filter parameters and the sequence length of input spectrogram for SELDnet was tuned using the development set of the respective subset. The performance of the tuned method was tested on the evaluation set of the subset, and the respective metrics averaged across the three cross-validation splits of the subset are reported.

Unlike the DNN-based method, the parametric method requires additional information on the number of active sources per frame to estimate the corresponding DOAs. However, SELDnet obtains this information from the data itself. In order to have a fair comparison, we used the minimum description length (MDL) [32] principle to estimate the number of sources from the input spectrogram and use it with MUSIC, resulting in the MUSIC output of MUS_{MDL} and the corresponding particle filter output of MUS^{PF}_{MDL}.

Finally, we studied the importance of recurrent layers for the SELDT task by removing them from SELDnet and evaluating the model containing only convolutional and dense layers, referred to as CNN hereafter. The best CNN architecture across datasets had five convolutional layers with 64 filters each.

4. RESULTS AND DISCUSSION

On tuning the input sequence length for SELDnet, it was observed that a sequence of 256 frames gave the best scores for the reverberant datasets, and 512 frames gave the best scores for the anechoic datasets. The SELDnet predictions and the corresponding references are visualized in Figure 2 for a respective 1000 frame test sequence from MANSYN $O2$ dataset. Each sound class is represented with a unique color across subplots. We see that the detected sound events are accurate in comparison to reference. The DOA predictions are seen to vary around the reference trajectory with a small deviation. This shows that SELDnet can successfully track and recognize multiple overlapping and moving sources.

Figure 3 visualizes the tracking predictions and their respective references for SELDnet and the baseline method MUS^{PF}_{GT}. In gen-

Table 2: Evaluation results on different datasets. Since the number of active sources information is used in $\mathrm{MUS_{GT}}$, the frame recall is always 100% and hence not reported. DE: DOA error, FR: Frame recall, F: F-score, SCOF: Same class overlapping frames

		ANSYN			RESYN			REAL			MANSYN			MREAL		
Tracking results		*O1*	*O2*	*O3*	*O1*	*O2*	*O3*	*O1*	*O2*	*O3*	*O1*	*O2*	*O3*	*O1*	*O2*	*O3*
$\mathrm{MUS_{GT}}$	DE	1.3	5.0	12.2	21.7	28.9	32.5	15.1	33.9	44.1	0.6	14.8	28.0	16.4	34.1	43.9
$\mathrm{MUS^{PF}_{GT}}$	DE	0.1	1.1	2.3	4.0	5.2	6.1	3.3	8.8	12.0	0.2	4.2	8.0	3.6	8.1	11.9
	FR	97.0	88.5	74.3	83.8	55.6	37.3	93.0	71.0	44.7	98.7	92.3	75.1	91.0	69.9	48.3
Methods estimating the number of active sources directly from input data																
$\mathrm{MUS_{MDL}}$	DE	0.5	14.2	24.0	22.3	31.9	38.5	25.3	36.2	44.1	4.2	17.8	28.5	26.5	35.9	44.9
	FR	93.9	**89.4**	**86.7**	61.7	45.6	52.5	53.6	35.7	**57.5**	63.8	48.1	51.85	53.4	35.2	**58.9**
$\mathrm{MUS^{PF}_{MDL}}$	DE	**0.1**	**4.4**	**7.2**	**6.4**	**10.6**	**12.7**	**9.3**	**10.9**	**13.7**	**3.5**	**6.8**	**8.0**	**13.6**	**11.2**	**13.6**
	FR	96.3	83.5	67.7	52.0	34.1	24.2	52.7	40.1	29.6	64	49.9	39.8	58.7	34.4	27.5
CNN	DE	25.7	25.2	26.9	39.1	35.1	31.4	32.0	34.9	37.1	26.1	25.8	28.2	36.6	39.3	40.2
	FR	80.2	45.6	32.2	69.5	45.8	29.7	45.1	28.4	16.9	83.7	58.1	38.3	44.5	26.2	16.3
SELDnet	DE	3.4	13.8	17.3	9.2	20.2	26.0	26.6	33.7	36.1	6.0	12.3	18.6	36.5	39.6	38.5
	FR	**99.4**	85.6	70.2	**95.8**	**74.9**	**56.4**	**64.9**	**41.5**	24.6	**98.5**	**94.6**	**80.7**	**69.6**	**42.8**	28.9
Detection results																
CNN	ER	0.52	0.46	0.51	0.44	0.45	0.54	0.52	0.51	0.51	0.59	0.47	0.48	0.46	0.49	0.52
	F	70.1	66.5	68	57	54.9	42.7	50.1	49.5	48.9	65.6	62.7	60.1	55.4	50.9	48.8
SELDnet	ER	**0.04**	**0.16**	**0.19**	**0.1**	**0.29**	**0.32**	**0.4**	**0.49**	**0.53**	**0.07**	**0.1**	**0.2**	**0.37**	**0.45**	**0.49**
	F	**97.7**	**89**	**85.6**	**92.5**	**79.6**	**76.5**	**60.3**	**53.1**	**51.1**	**95.3**	**93.2**	**87.4**	**64.4**	**56.4**	**52.3**
SCOF (in %)		0.0	4.2	12.1	0.0	4.2	12.1	0.0	7.6	23.0	0.0	3.0	9.1	0.0	7.1	20.9

eral, the performance of the two methods is visually comparable. Both methods are often confused in similar situations, for example in the intervals of 4-5 s, 10-13 s, and 23-25 s.

The SELDnet, by design, is restricted to recognize just one DOA for a given sound class. But in real life, there can be multiple instances of the same sound class occurring simultaneously. This is also seen in the datasets studied, the last row (SCOF) in the Table 2 presents the percentage of frames in which the same class is overlapping with itself. In comparison, the parametric method has no such restriction by design and can potentially perform better than SELDnet in these frames (even though, highly correlated sound events coming from different DOAs can easily degrade the performance of parametric methods such as MUSIC). The performance of the two methods in such a scenario can be observed in the 10-13 s interval of Figure 3. The SELDnet tracks only one of the two sources, while the parametric method tracks both overlapping sources and introduces an additional false track between the two trajectories.

Table 2 presents the quantitative results of the studied methods. The general trend is as follows. The higher the number of overlapping sources, the lower the tracking performance by both SELDnet and the parametric method. Across datasets, the DOA error improves considerably with the use of the temporal particle filter tracker, but at the cost of lower frame recall. By using MDL instead of reference information for the source number, the overall performance of the parametric approach reduces ($MUS^{PF}_{GT} > MUS^{PF}_{MDL}$). This reduction is especially observed in the frame recall metric, that drops significantly for reverberant and moving source scenario datasets, indicating the need for more robust source detection and counting schemes.

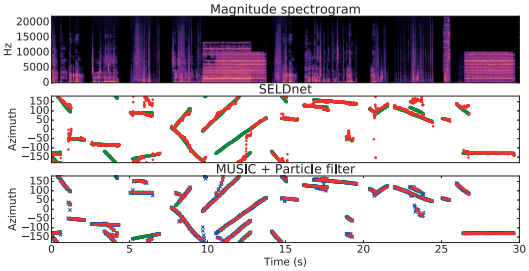


Figure 3: The tracking results of the two proposed methods are visualized for a MANSYN *O2* dataset recording. The top figure shows the input spectrogram. The center and bottom figures show the output of SELDnet and $\mathrm{MUS^{PF}_{GT}}$ tracker in red, and the groundtruth in green. The blue crosses in the bottom figure represents the frame-wise DOA output of MUSIC

The frame recall of SELDnet is observed to be consistently better than MUS^{PF}_{MDL}, but the DOA estimation is poorer across datasets. A similar relationship is observed between SELDnet and MUS^{PF}_{GT} for all the datasets generated with simulated impulse responses, while for the real-life impulse response datasets the frame recall of SELDnet is poorer than MUS^{PF}_{GT}. That could indicate the need for more extensive learning for real-life impulse response datasets, with larger datasets and stronger models.

Using recurrent layers definitely helps the SELDT task. It was observed from visualizations that the tracking performance by the CNN was poor, with spurious and high variance DOA tracks, thus resulting in poor DOA error across datasets as seen in Table 2. This suggests that the recurrent layers are crucial for SELDT task and perform a similar task as an RBMCDA particle filter of identifying the relevant frame-wise DOAs and associating these DOAs corresponding to the same sound class across time frames.

5. CONCLUSION

In this paper, we presented the first deep neural network based method, SELDnet, for the combined tasks of detecting the temporal onset and offset time for each sound event in a dynamic acoustic scene, localizing them in space and tracking their position when active, and finally recognizing the sound event class. The SELDnet performance was evaluated on five different datasets containing stationary and moving sources, anechoic and reverberant scenarios, and a different number of overlapping sources. It was shown that the recurrent layers employed by the SELDnet were crucial for the tracking performance. Further, the tracking performance of SELDnet was compared against a stand-alone parametric method based on multiple signal classification and particle filter. In general, the SELDnet tracking performance was comparable to the parametric method and achieved a higher frame recall for tracking but at a higher angular error.

6. REFERENCES

[1] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[2] ——, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016.

[3] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," in *Journal of Robotics and Mechatronics*, vol. 29, no. 1, 2017.

[4] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Int. Conf. on Robotics and Automation (ICRA)*, 2018.

[5] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *European Signal Processing Conference (EUSIPCO)*, 2011.

[6] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," in *IEEE Signal Processing Letters*, vol. 21, 2014.

[7] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," in *ACM Computing Surveys (CSUR)*, 2016.

[8] C. Grobler, C. Kruger, B. Silva, and G. Hancke, "Sound based localization and identification in industrial environments," in *IEEE Industrial Electronics Society (IECON)*, 2017.

[9] P. W. Wessels, J. V. Sande, and F. V. der Eerden, "Detection and localization of impulsive sound events for environmental noise assessment," in *The Journal of the Acoustical Society of America 141*, vol. 141, no. 5, 2017.

[10] R. Chakraborty and C. Nadeu, "Sound-model-based acoustic source localization using distributed microphone arrays," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[11] K. Lopatka, J. Kotus, and A. Czyzewsk, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools and Applications Journal*, vol. 75, no. 17, 2016.

[12] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*, 2015.

[13] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, 2018.

[14] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of Multiple Moving Speakers With Multiple Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[15] J. M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[16] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[17] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte Carlo data association," in *IEEE Workshop on Statistical Signal Processing (SSP)*, 2009.

[18] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1409–1415, 2012.

[19] J. Traa and P. Smaragdis, "Multiple speaker tracking with the Factorial von Mises-Fisher Filter," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014.

[20] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 392–402, 2014.

[21] J. Nix and V. Hohmann, "Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 995–1008, 2007.

[22] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 806–815, 2013.

[23] N. Strobel, S. Spors, and R. Rabenstein, "Joint Audio-Video Signal Processing for Object Localization and Tracking," in *Microphone Arrays*. Springer, 2001, pp. 203–225.

[24] J. Gu, X. Yang, S. De Mello, and J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] Y. J. Choe, J. Shin, and N. Spencer, "Probabilistic interpretations of recurrent neural networks," *Probabilistic Graphical Models*, 2017.

[26] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*. John Wiley & Sons, 2017, pp. 89–140.

[27] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, vol. 6, no. 6, 2016.

[28] H. W. Kuhn, "The hungarian method for the assignment problem," in *Naval Research Logistics Quarterly*, no. 2, 1955, p. 8397.

[29] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, 1986.

[30] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized particle filter for multiple target tracking," *Information Fusion*, vol. 8, no. 1, pp. 2–15, 2007.

[31] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference (EUSIPCO)*, 2018.

[32] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.